

AD 653722

FINAL REPORT

Conference on Cluster Analysis  
of Multivariate Data  
New Orleans, La., December 9, 10 and 11

This Conference was Sponsored by  
Office of Naval Research  
Contract No. N00014-67-C-0175

Principal Investigators  
Maurice Lorr & Samuel B. Iyerly

DDC

JUN 19 1967

Reproduction in whole or in part is permitted for any  
purpose of the United States Government

E

June, 1967  
Catholic University of America  
Washington, D. C.

ARCHIVE COPY

This document has been approved  
for public release and sale; its  
distribution is unlimited.

**SOME CRITICAL ISSUES AND PROBLEMS  
IN CLUSTER ANALYSIS**

Prepared for the  
Conference on Cluster Analysis of Multivariate Data  
December 9-11, 1966  
New Orleans, Louisiana

Samuel B. Lyerly  
BUREAU OF SOCIAL SCIENCE RESEARCH, INC.  
1200 Seventeenth Street, N. W.  
Washington, D. C. 20036

## SOME CRITICAL ISSUES AND PROBLEMS IN CLUSTER ANALYSIS

Samuel B. Lyerly

Although this paper is the first on the program, it is in no sense a "key-note address" or an introduction to the presentations that will follow. It is mainly an attempt to propose some questions that I and several others have been concerned about and for which we hope to get from this meeting some useful insights, if not definitive answers--if not from the papers, perhaps from the informal discussions for which we have budgeted a liberal proportion of time.

We all know that many problems in cluster analysis are common to various fields, including the several represented here; and if I lapse into the language of psychology from time to time I am sure that you will have no trouble making appropriate translations. And if some of my remarks seem critical of certain work that has been done, I am sure that you will understand that I am referring to others who are not present in this room.

I think it is well for us to remind ourselves that even in this enlightened decade "typological" concepts are controversial with many of our colleagues in the behavioral sciences. Back in my undergraduate days in psychology the prevailing doctrine was that individual differences are essentially quantitative rather than qualitative (and, if you used an appropriate measuring instrument and followed the instructions in the manual, they should all be "normally distributed"). Even unmistakably aberrant behavior, when it could not be linked to some physical injury or disease or to a genetic origin, was likely to be regarded as an extreme manifestation of some "normal" dimension of behavior. In recent years, however, there seems to have developed a growing suspicion that there may be ways of assigning people to groups or types or diagnostic categories in such a way that knowing a person's classification will significantly aid professional workers in helping him in medical, vocational, educational, or other situations. I am sure that if we did not share this point of view, or were not members of this "type," we would not be here today.

The first big question, and one which I am sure will receive a certain amount of attention during these several days, is: "What is a cluster?" For human populations I have seen no definition that can be unequivocally translated into operational procedures and few if any which seem to have satisfied even those investigators who have proposed and used them. A typical statement is that a cluster (type, group, species) is composed of individuals (objects, specimens, activities) such that every

member of the cluster is in some relevant sense "closer to" other members of his group than he is to members of other groups. As a definition, a statement such as this is of course circular and permits of various interpretations, depending upon the investigator's purpose, the nature of the data he happens to have at hand, and the computer program that he has been able to borrow. In a typical study in the social sciences one does not know at the outset whether any types or clusters exist (by whatever definition); how many clusters to expect, if any; what proportion of the sample can be comfortably assigned to one or another of the clusters that may be discovered; or what kind of statistical conclusion or probability statement can be made to reflect one's degree of confidence in the findings. There is often no preliminary statement of a clear model, either substantive or structural, that the investigator is seeking to confirm. In many psychological studies seeking clusters we may get useful hints about an implied theoretical model or about certain likely hypotheses by studying the list of variables the investigator has chosen to analyse. But this is not always a clear guide, since variables seem to be chosen frequently because of availability or for even more obscure reasons. So I hope to leave this conference feeling a little more secure about the cluster concept--what a cluster is, how to recognize one when I see one, what advances are being made toward operational, objective methods of cluster identification.

My second area of concern has to do with the choice of variables to be used in a cluster analysis. As I implied a moment ago, ideally the variables should be specified by the investigator's initial hypothesis or model, but in the typical "exploratory" study this is not always the case. Sometimes there does not appear to have been a clear understanding of the nature and characteristics of some of the variables employed. Occasionally sets of variables from quite dissimilar domains have been brought together in attempts to seek clusters within a common set of dimensions. Some research programs have taken what seems to me to be a sensible course (at least in those areas of psychology in which such a course is applicable): Variables are selected which are relevant in terms of the investigator's theory and whose characteristics or "meanings" are well understood from previous work (e.g., validity studies, factor analysis, or the like). I understand that in some fields, such as biological taxonomy, there are some fairly explicit models and that the selection of variables to be used in classification can thereby be more rationally determined. I hope that Professor Sokal will enlighten us on this.

Related to the selection of variables is the problem of their distributional form, and the associated problem of the metric properties of the data. Some investigators insist or prefer that only normal (or normalized) variables be used. Others do not hesitate to use nonnormal data, dichotomies, or orthogonal "dummy" components of multichotomous data. There is more than a matter of taste involved here. Are certain relevant data in the domain "inherently" nonnormal or qualitative? What are the scalar properties of a given variable? Considering the selective and/or haphazard conditions under which many of our human samples are drawn (in schools, hospitals, etc.) and the adventitious origins of many of the observations behavioral scientists use, how can we reasonably



expect or demand any particular distribution forms? One great advantage of normality (in particular, multivariate normality, which isn't easy to come by) is that it facilitates various statistical manipulations and permits certain significance tests. But cluster analysis is a long way from becoming a statistical method and in the meantime there are probably some more pressing problems deserving priority than the matter of whether distributions conform to the normal or any other standard shape.

One of the more critical problems in cluster analysis and related techniques is the choice of an interperson index, since the process usually starts with a table or matrix of  $n \times n$  numbers, each representing comparisons of each individual in a sample of  $n$  with every other individual. These indices, as you know, are typically one of two kinds: measures of similarity (correlations, covariances, cross-products, "per cent agreement") or measures of dissimilarity ("Euclidian" or some other index of "distance"). You are all familiar with these indices and their major characteristics. The point I want to make is that some investigators seem to have made their choice of index on the grounds of convenience or familiarity without recognizing that different indices can give rise to quite different cluster configurations. It is not necessary at this time or in this company to elaborate or document this statement. I shall be interested, however, to learn from some of our participants their reasons for choosing the indices they have used and their experiences and recommendations.

Incidentally, in a hasty and incomplete survey of the social science literature covering the past five or six years, I have found that the distance type of index is now leading the correlational type by about two to one. I think there may be several reasons for this: (1) Distance measures have received more respectful attention from statisticians, who have as you know developed some elaborate distance-based models for use in the closely related classification-decision problems. (2) Correlational indices ("Q" measures) have certain metric problems and seem to suffer from particular ambiguities from the sampling-significance point of view. (3) The use of the correlation coefficient involves the controversial "level" concept, which has not always been squarely faced. (My own feeling is that "level," which is an average, can be removed or ignored only when it is demonstrably irrelevant to the investigator's purpose and when it has a clear meaning in its own right, e.g., the mean or total score derived from a battery such as the Wechsler subtests. It follows, then, that the variables must be from the same domain, must all "point in the same direction" so far as their general behavioral significance is concerned, and hence be positively correlated.)

I shall pass over several related technical matters such as the appropriate dimensionality of one's space; whether the dimensions should be orthogonal or correlated; the questions of standardizing, weighting, etc., with the suggestion that perhaps we are not yet ready for decisions on some of them. Perhaps we need more experience with various empirical approaches which aspire no higher than the descriptive and the topological.

The area which has received the most attention recently, with the increasing availability of electronic computers, concerns the efficient

manipulation of data according to some routine or program designed to locate clusters if they exist and to assign each assignable member of the sample to his appropriate group.

With a matrix of interpersonal similarity or dissimilarity measures, there are two general methods of attack that have been used. The older, and still the most frequent, involves the locating of pairs of individuals who are "closest" to form the nuclei for types or clusters, then examining other individuals or pairs to be added to existing groups or to form tentative new groups. Various sequences or "rules" have been adopted and various criteria for inclusion, exclusion, or reassignment--some planned objectively (and hence adaptable to computer methods) and others dependent upon the investigator's judgment at various points in the process. The rules are essentially arbitrary and there are usually a number of individuals left unassigned to any group. This may be called the "synthetic" approach to the clustering process. (A British writer has recently called it the "agglomerative" method.)

The other major approach, which has been attempted more frequently in recent years, is what might be called the "analytic" method (or "divisive," in the term of our British colleague). Instead of beginning with  $n$  individuals, each a "cluster of order one," and successively combining pairs and larger groups until all or most have been assigned according to some rule of "belongingness," the investigator begins with the entire sample as one cluster and asks "How can I divide these into two groups, each of which is more homogeneous with respect to some criterion or standard than is the total sample and more homogeneous on the average than would be the case if any other partitioning into two groups were made?" The criterion may be something like minimizing within-groups sums of squares or maximizing between-groups differences.

Next, having divided the original sample into two groups according to the criterion (which, if carried out completely, involves examining each of the possible  $(2^n - 1)$  partitions), the investigator may analyze each of them and search for ways to divide them into further subgroups. This sequence of steps may be continued and the results tested at each stage (although an "exact" test of an appropriate null hypothesis for such a procedure is not known).

The result of this series of operations will ordinarily be an hierarchical "tree" configuration of groups, consisting of a "trunk" (the original undifferentiated sample), one or more orders of "limbs" and "branches," and finally the "twigs" (the ultimate smallest groups which cannot be further subdivided). The configuration need not be symmetric. Some ultimate categories may be at the limb or branch level.

Two characteristics of the "analytic" approach in comparison with the "synthetic" are: (1) it is more "objective" and hence more readily programmed for computers (at least in the forms in which recent investigators have used these methods, though not necessarily in general); and (2) it assures that every individual is assigned to one of the ultimate groups, provided some quasi-statistical criterion is used to terminate the

process (such as a predetermined within-groups sum of squares or a minimum number of cases in the ultimate categories).

Obviously, if either the synthetic or the analytic procedure is allowed to proceed unchecked by any rule of "when to stop," the Sorcerer's Apprentice will take charge. The synthetic approach will ultimately assign everyone to a single type, and the analytic will finally split the entire group into  $n$  classes, each containing one person.

Most of the attempts at empirical clustering have been step-wise and/or iterative procedures. A solution which has some obvious appeal is that the investigator form every possible arrangement and test each such arrangement against the criterion he has chosen. In other words, he would divide his subjects into every possible set of 2 groups, 3 groups, etc., and test every such set of partitions. This could be considered a frontal attack, avoiding some of the theoretical objections to the synthetic or analytic approaches. The difficulty with this idea is that with samples of even moderate size the problem is beyond the ability of even the fastest and most capacious modern computer to handle. The number of ways of classifying  $n$  individuals into  $r$  groups is  $n!/r!$  times the coefficient of  $x^n$  in the expansion of the generating function  $(e^x - 1)^r$ . For a sample of 16, which is certainly as small as most investigators would want to use, the total number of arrangements is more than 10 billion! Hence the need for short-cuts, approximations, and iterative approaches to the clustering problem.

In order to have more time for discussions, which we all hope will be a very fruitful part of this conference, I shall not continue along these lines at this time. My concluding summary remarks (and I have written some down) will be postponed until the end of the conference if anyone wants to hear them. I'll be very much interested in whether and to what extent I'll want to change them by that time.

## Methods of Cluster or Typological Analysis

Maurice Lorr

Catholic University

The purpose of this report is to review and examine available methods of typological or cluster analysis. To statisticians these techniques deal with what is known as the mixture problem. The mixture problem is concerned with a sample regarded as composed of individuals from several different populations. Neither the number of populations nor their nature are known a priori. It is also not known which individuals come from which populations. A general solution requires estimating both the number of populations present as well as the parameters of the different populations. Since the problem is very difficult, exceedingly little work of a probabilistic nature has been done.

Through usage, cluster analysis has come to refer to procedures applied for two different purposes. One reference is to procedures for identifying types, that is to say, homogeneous, mutually exclusive subsets of individuals, cases, objects or sampling units within a matrix of data. This process may be called typological analysis. In its second meaning, cluster analysis refers to procedures for grouping attributes, traits or characteristics. Here two different objectives may be distinguished. In one case the aim is data reduction or parsimony; a smaller set of measures are used to represent the larger set with a minimal loss of information. The second aim is to have each subset reflect some hypothetical dimension. The process may thus be called dimensional analysis. The concern here is with procedures for determining types not known a priori.

### The Utility of Typologies

What are some of the practical and scientific uses of typologies? It is obvious that a type facilitates communication. The unique pattern of type characteristics make members of a type easily recognized, remembered, understood and differentiated from non-members in a given domain. To label a person a psychopath or a schizoid immediately suggests a broad pattern of traits and to-be-expected behavior. A second related advantage is that type membership may provide enhanced predictions to outside criteria particularly if relations among variates are strongly nonlinear. A sample of persons of identical or homogeneous profile will tend to be more homogeneous as to criterion-relevant behavior than the mixed population (Toops, 1948). The integrity of the individual is preserved in the type concept since the entire score profile is considered simultaneously. Usually his scores are considered singly and in isolation. The improvement in predictive accuracy takes place through the operation of higher order dependencies and through the utilization of any interactions should they exist. In linear regression equations the predicted Y scores are simple weighted additive sums of the predictor scores in which the weights are constants. Interactive effects, like the simultaneous presence of say, two high scores and two low scores, are ignored. The possibilities of such configural relations have been shown by Mech1 (1940), Horst (1956), and by Lubin and Osburn (1957). For example, two dichotomous items may be totally unrelated to a dichotomous criterion (such as schizophrenic vs normal) when scored singly. Yet, when scored for their joint presence or absence, these two items may provide near perfect prediction to the criterion.

A taxonomy of natural occurring types represents an important achievement in

its own right. If there are discrete, qualitatively distinct subtypes present and demonstrable, then this knowledge reflects and increased understanding of the domain. The taxonomy may have much systematic import and generality. It may facilitate the discovery of laws not observable within mixed samples. The subgroups may provide or suggest information relative to common structure, common processes, and common antecedents much as they do in biology.

In opposition to the general purpose typing approach just described, the proponents of the single purpose approach argue that there is no single meaningful way to classify people. It all depends on one's purpose. Persons similar in one set of variables are not necessarily more similar than persons in general on another set of variables. A particular classification is meaningful only in so far as it is related to a broader class of variables one desires to predict or control. In this approach some mathematical function of the profile elements is found or constructed which will best predict the external criterion. Emphasis is on the criterion relevancy of the type characteristics and not on the nature of the profile. Finally it is argued that multiple linear or curvilinear regression is more efficient than prediction from knowledge of type membership.

It is true that there are numerous ways of classifying people in a given domain depending upon one's aims. However, the presumption in the mixture problem is that two or more natural subgroups exist. If they exist, they are likely to have arisen or developed because of survival value, or because of a conjunction of natural laws. In contrast the classification schemes and configural scores tied to external criteria represent technological advances lacking scientific generality. Each new decision and each particular situation calls for another empirical search for a criterion-relevant pattern. While useful for a while these cook-book patterns are soon outdated as new criteria or potential predictors appear. The argument against special-purpose types is comparable to that offered in support of the development of psychological tests as instruments of psychological theory (Loevinger, 1957; Cattell, 1946). Just as criterion-oriented psychometrics and particularized validation are devoid of scientific interest, so are single-purpose classification schemes.

### Structural Models

Before examining specific procedures for finding subsets of entities the problem of structural models requires consideration. The overall problem is one of developing a fruitful means for representing the data. Cluster-search procedures should determine rather than impose structure on a body of data. If, for example points are uniformly distributed in space no clusters should be found. Indeed empirical data suggest that clusters may vary greatly in shape. In three-dimensional space, they may be spheroid, serpentine, amoeboid or cloud-like. Thus it should be evident that quite different cluster-search methods are needed to ascertain different structures and different objectives. There should be no arbitrary partitioning or chopping up of space.

### Cluster-Search Techniques

The cluster-search procedures may be classified for purposes of description and discussion into the following categories: (a) factor analysis, (b) multi-dimensional scaling (c) minimizing within-cluster variation, (d) successive cluster build-up, (e) linkage analysis and (f) hierarchical analysis.

### A. The Method of Factor Analysis

Factor analysis of the N by N matrix of interperson similarities followed by a rotation to simple structure has been a common procedure for identifying types. Stephenson (1936), Tryon (1955), Bass (1957), Broverman (1961), Nunnally (1962), Overall (1964) and many others have recommended this procedure. The indices of resemblance may be correlations, normalized crossproducts of scores, squared distances, or simply crossproducts of raw scores.

Factor analysis is deemed inappropriate because the method is designed to isolate dimensions and not clusters of entities. There is no reason why clusters defined by two or more dimensions may not be more numerous than dimensions. The rotational process also is inappropriate for the task of isolating mutually exclusive subgroups. The usual rotational process tends to dismember clusters or to miss them altogether. If a cluster should happen to fall between two factors, each type-factor will be defined by persons on the margins of the cluster. Also factoring tends to yield a multiple classification of persons since most persons will correlate significantly with several type-factors, and relatively few with one factor.

When correlations, covariances and normalized crossproducts are factored, all unrotated factors are bipolar and such bipolarity cannot be completely removed (Ross, 1963). Thus persons with opposite score profiles will emerge with high but opposite loadings on the same factor. Each type-factor is, therefore, defining two types rather than one. Thus, the number of type-factors defined cannot be the same as the number of types.

The most cogent general argument advanced against the use of factor analysis of similarity indices between persons is that it does not yield new information. The number of factors resulting from a direct R-analysis of measures and an obverse Q-analysis of persons will be the same (Burt, 1937; Harris, 1955; Slater, 1958; Ross, 1963; Ryder, 1964). If variables have been standardized over subjects, a principal component analysis of sums of score profile crossproducts yields exactly the same results as an analysis of correlations among variables.

Lazarfeld's latent class model (1950) as further extended by Gibson (1959) also calls for a factor analysis. The technique operates on the interrelations of dichotomous attributes. Manifest joint frequencies are accounted for by a set of Q mutually exclusive and exhaustive subgroups (latent classes). The model assumes that each subgroup or latent class is homogeneous in whatever underlying dimensions are necessary to account for the observed interrelations. Stated otherwise, there is within-class independence between pairs of tests. The number of latent classes is determined by a factor analysis of the lower order joint occurrence matrix.

One question that can be raised is how configural information from higher order joint occurrences can affect this solution. Lunnenborg (1959) has argued that the independence of items effectively precludes the possibility of configural information unless the latter is present in the sets of items prior to the determination of latent classes. Another limitation to the method is that it appears to be confined to variates of relatively small dimensionality--usually one or two. Most typing problems in psychology involve at least six or more dimensions.



Although there is nothing in the development of the model equations that restricts the number of dimensions, empirical examples involving more seem not to have been published. Other obstacles are that latent class sizes must be estimated or known in advance.

## B. Minimizing Within Cluster Variation

One procedure, often proposed, is to subdivide the  $N$  profiles in  $K$ -space into  $Q$  mutually exclusive subsets in such a way that each is as compact and homogeneous as possible. Compactness is achieved by requiring the average of all distances between profiles within each subset shall be a minimum. This technique has been variously labeled a "minimum variance partition" and a "minimum squared error technique."

One of the first of such efforts was reported by Thorndike (1953). His procedure begins by assuming that the two profiles which are the greatest distance apart fall into different subgroups. A third subgroup is established with a profile which is furthest away from either of the other two. Each cluster is built up by adding that profile nearest the pivot defining the cluster. A profile is added to each cluster in turn until all specimens are assigned. This yields sets of clusters of equal size. Profiles found closer to members of another cluster than to their own are re-assigned until further shifts do not reduce within-cluster distances. Increases in the number of clusters are made in the same manner until the average within-cluster distances relative to the number of clusters stabilize. While the procedure is comparatively objective it has some limitations, a few of which will be mentioned. For instance the goal of assigning specimens so that the average within-cluster distances are at a minimum involves a fair degree of trial and error and no criterion for optimal termination. There are no limits set in assigning profiles close to two clusters; every profile is allocated to a cluster. There also appears to be no justification for assigning every profile to a cluster, not for seeking subgroups of equal size. Finally the number of groups must be specified in advanced.

Zubin, Fleiss, and Burdock (1963) have proposed a procedure for fractionating a population into homogeneous subgroups that resembles Thorndike's. First the matrix of  $D^2$ 's is scanned and the largest entry identified. The two profiles involved, say  $X$  and  $Y$ , then form the foci of two subgroups. About each of these foci separately is clustered each profile whose  $D^2$  from the focus is less than the fifth centile of all the squared distances. These two clusters are taken as nuclei. Then about each of these nuclei are clustered profiles whose average  $D^2$  from members of the nucleus is less than the tenth centile of all distances. The criterion of inclusion may be relaxed still further until every profile in the sample has been assigned to one of the subgroups. A profile that satisfies a criterion for both clusters is assigned to the group to which it is closer. The subgroups are then tested by chi square for homogeneity. If the clusters are not yet homogeneous, the next step is to identify that trio of profiles mutually furthest apart from one another than any other triplet. Profiles are again clustered about each of these foci and the homogeneity of the resulting subgroups is tested. This procedure is continued either until all groups are homogeneous or the number of groups to be found is so great as to be meaningless. The procedure assumes normality in the underlying groups, independent measures, and equal covariance matrices. The method tends to guard against the detection of

spurious clusters since it allows for the possibility that the population studied is homogeneous to begin with.

Forgy (1965) has delineated some of the shortcomings of the minimum variation technique. As an illustration he cites data from the field of astronomy reported by Hertzsprung and Russell. When stars are plotted by absolute luminosity and temperature two "natural" groups of stars are evident. The so called "main sequence" stars appear as a flat S pattern while the "red giants" group together in a compact cluster. A minimum variance partition of such a sample could cut right across these groups since such a partition would produce a smaller within-group sum of squares. Thus the method tends toward the arbitrary partitioning of space into "efficient" subsets. It is unsuited for the recovery of natural subgroups differing in configuration.

### C. Successive Cluster Buildup

In this technique either a single pair of profiles (usually the closest pair) or a profile with greatest variance is selected as a nucleus for the cluster. Other profiles are assigned to the cluster on the basis of a definition of similarity which sets a limit or threshold for inclusion. The method does not need to specify the number of clusters to be determined in advance.

McQuitty (1961, 1963) has developed several procedures, called typal analysis, representative of successive cluster buildup. He defines a type as a category of  $N$  people such that everyone in the category is more like each of the other  $N-1$  persons than he is like any other person in any other category. The method starts with a table of similarity indices between people. The indices of every column are then arranged in rank order and submatrices are built that satisfy the definitions of type. A submatrix satisfying the definition of type contains no rank larger than the number of persons in the type. Suppose a type consists of persons A and B, A being most like A and second most like B, and B in turn being most like B and second most like A. Then the submatrix constitutes a type if it contains no rank larger than the number of cases. This process continues until all persons of the original matrix have been chosen in order of their similarity to A. The problem is to select from the full matrix of indices all of the submatrices which fulfill the definition of a type. The advantages claimed for the method are that (a) it can reject an hypothesis of types; (b) it reports exceptions to a type. If typal analysis fails to yield types it is possible to relax the definition and permit inclusion of persons with slightly higher ranks than are permitted by the usual definition.

Sawrey, Keller, and Conger (1960) also have designed a cluster buildup procedure which uses the distances ( $D^2$ 's) between each and every profile. First an arbitrary maximum  $D^2$  is set as a definition of "similarity." Then with each profile are listed all other profiles in the matrix whose distance is less than the maximum. The profile with the largest number of other profiles similar to it is selected to form a potential nucleus group. The profile selected and all those similar to it are crossed out from the table. The profile with the next highest number of similar profiles is then selected to become the second potential nucleus group. Again the associated list of profiles is crossed out from the table. The process is repeated until only profiles having no similar profiles remain. Next a minimum value is set for the definition of "dissimilarity" and a matrix of the



selected profile indices is prepared. The columns of the matrix are summed and dissimilar pivot profiles are selected. Selection proceeds from the profile having the largest sum to the profile having the smallest sum. As a profile is selected all other profiles which are not dissimilar to it (i.e., whose distance from the selected profile is less than the maximum) are eliminated from the matrix. The selected profiles are all at least the minimum distance from each other. The centroid of each nucleus group (the selected profile and associated list) is determined. Each remaining profile is added to a nucleus group if its distance is less than the limit of dissimilarity from any member of the nucleus group. Several maximum limits may be set for adding in additional profiles to existing groups. Only an upper limit is used to form the nucleus groups. Although distances among members of a cluster may vary greatly, these are ignored. Several maxima would appear needed to define similarity since a subgroup whose members are more widely separated from each other and from other groups will remain unrecognized.

Saunders and Schucman (1962) have developed a procedure, called syndrome analysis, that satisfies McQuitty's definition of type but operates on squared distances between profiles. It begins by regarding every individual in the sample as a cluster of order one. First, all pairs that are mutually closest to each other are identified. Then all triplets whose members are closest to each other are found. Clusters of higher order are identified by the same process until no more clusters appear by this process. A list of "closed clusters" is examined to eliminate those which are contained in larger closed clusters that came to light later in the process. The resulting list of non-overlapping closed clusters are regarded as "nodes" for the given matrix. The third step is to characterize the nodes. This may involve finding the mean profile of members of each node, or it may involve construction of the within-node-variance-covariance matrix of test scores. The latent roots and vectors of the matrix may provide the necessary coefficients for partialling out intra-node variability preparatory to iteration of the procedure. Once membership has been established the resulting subset is called a syndrome.

Several cluster-search procedures similar to those just described have been developed by Lorr and his associates (Lorr, et al, 1962; Lorr and Radhakrishnan, 1967). The procedure begins by finding a profile near the center of a cluster. The profile with the maximum variance of squared correlations (or congruency coefficients) with all others is selected as pivot. To the pivot are added successively the two profiles with the highest average correlation with all profiles correlating above  $C_L$  with the pivot. The limit  $C_L$  may be set at the value at which a correlation coefficient based on  $K$  independent variates is significant at  $p$  less than .05. The matrix is searched and the profile added that correlates highest on the average with those already in the cluster. The process continues until no other profiles can be found that correlate on the average above  $C_L$ . Next an upper limit  $C_U$  is set to define dissimilarity and to prevent cluster overlap. A suitable value is a correlation coefficient significant at  $p$  less than .10. Any coefficient in the residual matrix that correlates on the average  $C_U$  or higher with the first cluster is deleted. The second cluster is generated in the same manner as the first from the matrix of remaining profiles. The deletion of profiles correlating above  $C_U$  with a newly formed cluster does not exclude profiles correlating above  $C_U$  with preceeding clusters. Accordingly, cluster members that correlate on the average above  $C_U$  with the last generated cluster are also deleted.

The final steps consist in determining (a) the mean correlations within and between clusters; (b) the mean standard score profile of each cluster. The computer program can handle 150 profiles at one time.

Like McQuitty's typal analysis, the procedure proposed by Gengerelli (1963) is based on a definition of a subgroup. Consider a population of  $N$  persons each measured on  $K$  variates. Let each person be represented as a point in  $K$ -dimensional space. Then a subgroup is defined as an aggregate of points in the test space such that the distance between any two points in the set is less than the distance between any point in the set and any point outside of it. Suppose  $N$  persons as points are distributed in three-dimensional space as two spheres, A and B. Two subsets will exist only if the two spheres are separated by a distance greater than the diameter of the larger sphere. The method begins with an  $N$  by  $N$  matrix of squared distances. A frequency distribution is made of distances between all possible pairs. The existence of one or more discontinuities in the distribution of distances indicates that a population consists of two or more subsets. The first point of discontinuity in the distributions,  $D_c$ , provides a criterion for determining the point of separation between two subsets. A subset is then defined as the aggregate of points (persons) who are mutually no farther apart one from another than  $D_c$ . The existence of subsets in a population is thus associated with multimodality in the distribution of inter-point distances. Computer programs and empirical tests are as yet not available.

Bonner (1964) has been responsible for several programs for clustering binary attributes, one of which has been generalized to continuous data (Pettit, 1964). One program is based on a type definition resembling McQuitty's. The goal is to find clusters where all members are similar to each other and no non-member is similar to all members. The algorithm picks a random "center" and builds a cluster around this through use of an arbitrary threshold  $T$ . Profiles more similar to the center than  $T$  are considered to be in the crude cluster. The typical member of the cluster is computed and compared with the expected number of clusters rarer than this to be found in an uncorrelated population. Then by means of a process of "hill climbing" a better cluster is achieved. All profiles are used as cluster centers.

Rogers and Tanimoto (1960) have reported a computer program for the classification of plants. Their variables are binary and a simple similarity coefficient is used. After a matrix of similarity coefficients has been obtained a value  $R_j$  is computed as a measure of the number of nonzero similarity coefficients possessed by a given individual. Next computed is a quantity  $H_j$  which is the product of all the similarity coefficients of  $j$  with others. All persons are then grouped in a table in order of descending value of  $R_j$ . The person having the highest  $R_j$  and the highest  $H_j$  is considered the prime mode. The problem is to find a criterion to determine the number of persons who go into a cluster. To do this a second node is found. The radius around the first node must be such as not to include the second node. At this point the similarity coefficients are converted into distances defined as  $D_{ij}$  equals  $-\log_2 S_{ij}$ . These distances permit visualization of taxonomic similarity. Finally a measure of cluster inhomogeneity is computed. The method has proved to be fairly effective in isolating subsets when the variables are truly qualitative categories.

Cattell and Coulter (1966) have developed a procedure that represents a

variant of cluster buildup. Given a matrix of similarity indices the next step is to establish several arbitrary limits as definitions of similarity. The matrix of similarities is then converted into an "incidence" matrix of ones and zeros. If an index exceeds the limit it is categorized as a unit to designate a linkage; otherwise it is categorized as a zero. Next a "phenomenal cluster" is defined as a set of profiles each of which is linked to every other. Spatially this means that all points fall within a hypersphere. A Boolean algorithm, based on what has been called "ramifying linkage method", sorts the data into phenomenal clusters.

#### D. Linkage Analysis

Linkage analysis classifies profiles into clusters such that every profile in a cluster is more like some other profile in that cluster than it is like any other profile in any other cluster (McQuitty, 1957). This method is especially useful in determining elongated, serpentine or amoeboid clusters. Profiles are continuously connected with one another through intermediate profiles thus maintaining any specified level of similarity. Linkage analysis has also been much applied to generate hierarchies which will be considered later.

McQuitty (1957, 1964) has been among the first to develop linkage analysis which is perhaps the simplest of the cluster methods. The analysis starts with a matrix of similarity indices. First the highest entry in each column (a linkage) is found, and then the highest entry in the matrix is identified. The highest entry (ab) represents a reciprocal pair in the sense that members are mutually closest to each other. One member of the pair (b) may also be the highest entry in some other columns, say c and d. Then c and d also constitute members of the cluster. If none of the profiles, a, b, c and d is highest in any other column the cluster is complete. The highest remaining entry in the matrix is then used to build the next cluster. Analogously, additional clusters are determined.

Cattell and Coulter (1966) also employ a procedure akin to linkage analysis to identify strung-out clusters. Instead of beginning with individual profiles they first identify all possible phenomenal clusters, (hyperspheres). The amount of overlap of the phenomenal clusters is recorded in a matrix which is then converted, through application of a limit, into an incidence matrix of units and zeros. This latter matrix is subjected to their search procedure which identifies all mutually exclusive chains of continuously related profiles.

Needham (1961) and Parker-Rhodes (1961) use linkage analysis with binary data. The distance between all pairs of profiles is determined. A limit or cutting score is set to define similarity and applied to the matrix which is reduced to a matrix of zeros and ones. Columns of the matrix are then compared pair-wise to determine the number of agreements or intersections between them. The resulting subsets, called "clumps", are defined as members more like each other and less like non-members than numbers of the universe picked at random.

The method of single linkage has also been suggested by Sneath (1957) and applied to taxonomic problems in biology (Sokal and Sneath, 1963). They point out that in avoiding overlapping clusters data may, in fact, be distorted to yield discrete clusters. When single linkages are permitted then complicated serpentine clusters may be formed. More will be said under the topic of hierarchical

clusters.

#### E. Multidimensional Scaling

Metric and nonmetric multidimensional scaling represents another possible as yet untried approach to cluster identification. Given an  $N$  by  $N$  matrix of interperson similarities some of the standard routines developed by Torgerson (1958), Shepherd (1962), Kruskal (1964) and Lingoes (1965, 1966) could be applied. In these procedures, individual profiles would be treated as points in space of unknown dimensionality. The problem would be to determine the dimensionality of the space and the location of the points in space. In the final solution distances between points in the space correspond to some monotonic function of the similarity of the corresponding profiles. The Guttman-Lingoes procedures are designed for the treatment of categorical qualitative data but are also adapted for use with quantitative data.

#### F. Hierarchic Cluster Analysis

Discussion, thus far, has been restricted to techniques for finding unordered qualitative classes or so-called natural clusters. Some of the procedures described have also been applied or extended to the problem of establishing discrete clusters each subdivided into subclasses. While there is some question in regard to the range of application of such methods to psychological problems, their use in biology is widespread. Sokal and Sneath (1963) assert that biological classification should be constructed by nested overlapping categories (p. 192). Thus some of the procedures for constructing hierarchic structures will be reviewed briefly.

Sneath's (1957) single linkage procedure is followed for the first sets. Then the criteria of admission (thresholds) are gradually lowered from an initial high similarity value to low similarity values. Thus a single link between any member of two clusters permits the establishment of a more inclusive cluster.

McQuitty has been responsible for numerous procedures for hierarchical classification (1954, 1960, 1964). Agreement analysis classifies objects into successive levels such as species, genera, and families. The first species is the two objects with the highest agreement score, the second species are the two objects with the next highest score. Species are then classified into more inclusive groups analogous to the way in which individuals were classified. Hierarchical linkage analysis seeks to classify individuals into categories such that every member of every category has a maximal number of common characteristics and a minimal number of categories are required. Later modifications have led to what is called hierarchical classification by reciprocal pairs and by typal analysis.

Ward (1963) and Ward and Hook (1963) have developed a very efficient minimum-within-group distance procedure for hierarchical grouping of profiles. Each larger group is a unique combination of the next subordinate subgroup. The technique operates on an  $N$  by  $N$  matrix of profile distances. Clusters are built up by adding cases which increase the mean within squared distance least. Clustering starts with  $N$  groups of one and ends with one group of  $N$ . Initially the matrix is scanned to find the pair of profiles with the smallest distance, these are combined to form a cluster of two. The distances of the remaining profiles from this cluster centroid are then computed. The process continues by reducing

the number of clusters from  $N$  (the original number) to  $N-1$ ,  $N-2$ ,...etc. at each stage the within-groups sums of squares is minimized. In addition they utilize an "objective function" which reflects the investigators purpose to guide the process.

The Ward technique assumes nothing about the underlying structure of the groups or their distributions. In fact it can partition any collection of profiles whether or not it contains "natural" groups. The multivariate distribution may even be multivariate normal and thus unimodal. They offer no statistical test as to how many groups are present. It is also likely that the nature of the groups established may depend on chance variations in data. Many similar comments can be made relative to the McQuitty techniques although they tend to be set-theoretic in form. On the other hand it can be argued that these procedures are in fact quite useful. They group jobs so as to reduce cross-training time, they facilitate retrieval of information, and they increase predictive efficiency.

Edwards and Cavalli-Sforza (1965) also apply the minimum-within-cluster sums of squares technique to construct hierarchic arrangements of clusters. The profiles are divided into the two most compact clusters, and the process is repeated sequentially so that a tree diagram is formed. The advantage of a tree representation is that it can be mapped on paper in two dimensions. Beginning at the base of the tree the first bifurcation represents the first split of profiles into two clusters. Each branch is split again as the two clusters are resolved into two more, and the process continued until individual points are reached.

### References

- Ball, G. H. Data analysis in the social sciences: What about the details? Proc. of the Fall Joint Computer Conference. 1965, 533-559.
- Ball, G. H. & Hall, D. J. ISODATA, a novel method of data analysis and pattern classification. Stanford Research Institute, Menlo Park, Calif. (Apr., 1965).
- Bass, B. M. Iterative inverse factor analysis: A rapid method for clustering persons. Psychometrika, 1957, 22, 105.
- Bonner, R. E. On some clustering techniques. IBM Journal of Res. and Dev. 1964, 22-33.
- Broverman, D. M. Effects of score transformations in Q and R factor analysis techniques. Psychological Review, 1961, 68, 68-80.
- Burt, D. Correlations between persons. Brit. J. Psych. 1937, 28, 167-85.
- Cattell, R. A note on correlation clusters and cluster search methods. Psychometrika, 9, (3) (sept., 1944).
- Cattell, R. B. The three basic factor-analytic designs--their interrelations and derivatives. Psychological Bulletin, 1952, 49, 499-520.
- Clement, F. E. The use of cluster analysis with anthropological data. American Anthropologist, 1954, 56, 180-199.
- Cox, D. R. Note on grouping. J. Amer. Stat. Ass., 1957, 52, 543-547.
- Driver, H. E. Survey of numerical classification in anthropology. Part Two, V, 301-344; in Hymes, D. (ed.), The use of computers in anthropology.
- Fisher, R. A. The use of multiple measurements in taxonomic problems. Am. Eugen., 1963, 7, 179-188.
- Fisher, W. D. On grouping for maximum homogeneity. J. Amer. Stat. Ass. 1958, 53, 789-798.
- Forgy, E. W. Detecting "natural" clusters of individuals. Paper read at Western Psychol. Ass., Santa Monica, Calif., April, 1963.
- Forgy, E. W. Evaluation of several methods for detecting sample mixtures from different n-dimensional populations Abstract. American Psychological Association Meetings. Los Angeles, September, 1964.
- Fortier, J. J. & Soloman, H. Clustering procedures. Tech. Report 7, Dept. of Statistics, Stanford University (March 20, 1964).
- Gengerelli, J. A. A method for detecting subgroups in a population and specifying their membership. J. Psychol., 1963, 55, 457-468.

- Gerard, R. W. & Mattsson, N. The classification of schizophrenia. Jacquiz, J. A. (ed.) The diagnostic process. Ann Arbor: Mallory Lithograph Inc., 1964.
- Gibson, W. A. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. Psychometrika, 1959, 24, 229-252.
- Green, B. F., Jr. A general solution for the latent class model of latent structure analysis. Psychometrika, 1951, 16, 151-166.
- Harris, C. C. A scientific method of districting. Behavioral Science, 1964, 9, July.
- Harris, C. W. Relations among factors of raw, deviation, and double-centered score matrices. J. Exp. Educ. 1953, 22, 53-58.
- Harris, C. W. Characteristics of two measures of profile similarity. Psychometrika, 1955, 20, 289-297.
- Helmstadter, G. C. An empirical comparison of methods for estimating profile similarity. Educ. & Psychol. Meas. 1957, 17, 71-82.
- Holzinger, K. J. & Harman, H. H. Factor analysis. Chicago: University of Chicago Press, 1941.
- Hymes, D. (ed.). The use of computers in anthropology. London: Mouton & Co., 1965.
- Hyvarinen, L. Classification of qualitative data, British Info. Theory J., 1962, 83-89.
- Ihm, P. Automatic classification in anthropology. Part Two, V, 357-378; in Hymes, D. (ed.), The use of computers in anthropology.
- Jones, K. J. The multivariate statistical analyzer. Cambridge, Mass., Harvard Cooperating Society, 1964.
- Kaskey, G. et al, Cluster formation and diagnostic significance in psychiatric symptom evaluation. Proc. Fall Jt. Computer Conf., 1962, 285.
- Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 1964a, 29, 1-28.
- Kruskal, J. B. Non-metric multidimensional scaling: A numerical method. Psychometrika, 1964b, 29, 115-130.
- Lazarfeld, P. F. Chapters 10 and 11 in Stouffer, S. A. et al. Measurement and prediction. Princeton: Princeton University Press, 1950.
- Lingoes, J. C. Multiple scalogram analysis: A set-theoretic model for analyzing dichotomous items. Educational and Psychological Measurement, 1963, 23, 501-24.



- Lingoes, J. C. A taxonomic optimization procedure: An IBM 7090 classification program. Behav. Sci., 1963, 8, 370.
- Lingoes, J. C. An IBM 7090 program for Guttman Lingoes smallest space analysis. Behav. Sci., 1966, 11
- Loevinger, Jane, Gleser, Goldine C. & DuBois, P. H. Maximizing the discriminating power of a multiple-score test. Psychometrika, 1953, 18, 309-317.
- Lunnenborg, C. E., Jr. Dimensional analysis, latent structure and the problem of patterns. Seattle: University of Washington, 1959.
- Lykken, D. T. A method of actuarial pattern analysis. Psychol. Bull. 1956, 53, 102-108.
- MacNaughton-Smith, P., Williams, W. T., Dale, M. B. & Mockett, C. G. Dissimilarity analysis: A new technique of hierarchical subdivision. Nature, 1964, 202, 1033-1034.
- McQuitty, L. L. Agreement analysis: Classifying persons by predominant patterns of responses. Brit. J. Psychol. Stat. Sect., 1956, 16, 68-73.
- McQuitty, L. L. Hierarchical Linkage Analysis for the isolation of types. Ed. Psych. Measmt., 1960, 20(1).
- McQuitty, L. L. Typal analysis. Ed. Psych. Measmt., 1961, 21, 677-696.
- McQuitty, L. L. Rank order typal analysis. Ed. Psych. Measmt., 1963, 23(1).
- McQuitty, L. L. Capabilities and improvements of linkage analysis as a clustering method. Ed. Psych. Measmt., 1964, 24, 441-456.
- McQuitty, L. L. Similarity analysis by reciprocal pairs for discrete and continuous data. Ed. Psych. Measmt., 1966, 26, 825-831.
- Meehl, P. E. Configural scoring. J. consult. Psychol., XIV (1950), 165-171.
- Michener, C. D. & Sokal, R. R. A quantitative approach to a problem in classification. Evolution, 11, 130-162, (June, 1957).
- Needham, R. M. Computer methods for classification and grouping. Part Two, V, 345-356; in Hymes, D. (ed.), The use of computers in anthropology.
- Needham, R. M. The theory of clumps, II. Report M. L. 139, Cambridge Language Research Unit, Cambridge, Eng. (Mar. 1961).
- Nunnally, J. The analysis of profile data. Psych. Bull., 1962, 59, 311-319.
- Orr, D. B. A new method for clustering jobs. J. Appl. Psychol., 1960, 44, 44-49.
- Parker-Rhodes, A. F. Contributions to the theory of clumps. I. M. L. 138, Cambridge Language Research Unit, Cambridge, Eng. (Mar. 1961).



- Parker-Rhodes, A. F. & Needham, R. M. The theory of clumps. Cambridge Language Research Unit, 1960.
- Peiloff, R. Persons selection: A technique for grouping a minimum number of persons to maximally predict a person-prototype. Brit. J. of Stat. 1963, 41, 211-213.
- Pettit, R. G. Clustering program continuous variables. Advanced Systems Developing Division, IBM, Yorktown Heights, N.Y., 1964.
- Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. Science, 1960, 132, 1115-1122.
- Ross, J. The relation between test and person factors. Psychol. Rev., LXX (1963), 432-443.
- Ryder, R. G. Profile factor analysis and variable factor analysis. Psychological Reports, 1964, 15, 119-27.
- Sakoda, J. M. Osgood and Suci's measure of pattern similarity and Q-technique factor analysis. Psychom., 1954, 19, 253-256.
- Saunders, D. R. & Schucman, H. Syndrome analysis: An efficient procedure for isolating meaningful subgroups in a nonrandom sample of a population. Paper read at 3rd Annual Psychonomic Society Meeting, St. Louis, Missouri, Sept., 1962.
- Sawrey, W. L., Keller, L., & Conger, J. J. An objective method of grouping profiles by distance functions and its relation to factor analysis. Ed. Psych. Measmt., 1960, 20, 651-674.
- Shepard, R. N. The analysis of proximities: Multi-dimensional scaling with an unknown distance functions I and II. Psychometrika, 1962a, b, 27, 125-140; 219-246.
- Slater, P. The general relationship between test factors and person factors. Nature, 1958, 181, 1225-1226.
- Sneath, P. H. A. & Sokal, R. R. Principles of numerical taxonomy. San Francisco: Freeman, 1963.
- Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, Mar. 20, 1958.
- Stephenson, W. Some observations on Q technique. Psychol Bull., XLIX (1952), 483-498.
- Thorndike, R. L. Who belongs in the family? Psychometrika, 1953, 18, 267-276.
- Tiedman, D. V. On the study of types. in Sells, S. (ed.) Symposium on pattern analysis. Randolph Field, Texas: USAF School Aviation Medicine, 1955.
- Toops, H. A. The use of addends in experimental control, social census, and managerial research. Psychological Bulletin, XLV (1948), 41-74.

Tryon, B. Cluster analysis. Ann Arbor: Edwards Brothers, 1939.

Tryon, R. C. Identification of social areas by cluster analysis. Berkley, Calif., 1955.

Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. 1963, 58, 236-244.

Wolfe, J. H. A computer program for the maximum likelihood analysis of types. Tech. Bull., 65-15, U. S. Naval Personnel Research Activity, San Diego, Calif., 92152 (May, 1965).

Zubin, J., Fleiss, J. & Burdock, E. I. A method for fractionating a population into homogeneous subgroups. Unpublished paper, 1963.

## A review of clustering methods in biological taxonomy<sup>1</sup>

Robert R. Sokal

The University of Kansas

### Introduction

If I interpret my task this morning correctly, it is to present to an audience composed largely of psychologists and other social scientists a review of the clustering methods which biological taxonomists have employed in recent years. There has been considerable activity in this field which, as many of you know, has come to be called numerical taxonomy. Although we biologists are newcomers in this field compared to the social scientists we have managed to accumulate a variety of methods in relatively few years. So, I could, in fact, report on a substantial number of different clustering approaches. However, I shall only sketch in scant outlines, since with one or two minor exceptions the techniques in biology are fundamentally akin to those of the social sciences (Ball, 1965), and there seems little point in reintroducing you to methods long familiar but disguised in biological garb.

You have undoubtedly been struck by the wide generality of your approaches across other disciplines of science. But it is important not to be overly impressed by this phenomenon. There are, in fact, fundamental differences, not in the mechanics of cluster analysis, but in the philosophical assumptions accompanying its use, which differ markedly among the various fields of application. And I hope to spend the greater part of my time explaining to you the bases of these assumptions in biology to permit you to contrast these with the assumptions upon which you have been basing your work. I feel that such an approach should be of interest to you. Through an appreciation of the differences in approach in clustering philosophy in other sciences I have gained more insight into my own research field and possibly similar benefits may accrue to you from such a comparative approach.

### Principles of taxonomy

Before we proceed we should define taxon as meaning a taxonomic group or class of any nature and rank. Operational taxonomic units (OTU's) are the lowest ranking taxa in a given study. They are the basic units that are to be grouped into higher ranking taxa. A character is a property or feature which varies from one OTU to another. It is coded into distinguishable states. Thus hairiness of a leaf is a character. Slight, medium and

---

<sup>1</sup>Contribution No. 1347 from the Department of Entomology, The University of Kansas. Based upon research supported by grant (AI-04438-03) from the National Institutes of Health and by a Public Health Service research career program award (3-K3-GM-22, 021-01S1) from the National Institute of General Medical Sciences.

heavy may be the three states in which this character occurs among the OTU's to be classified. A possible source of confusion is the difference between the terms "classification" and "identification." Where a set of unordered objects has been grouped on the basis of like properties, biologists call this "classification." Once a classification has been established, the allocation of additional unidentified objects to the correct class is generally known as "identification." Some mathematicians and philosophers would also call this second process "classification," but I am principally concerned with classification in the biologist's sense.

Fundamental criteria of a classification have been defined by Williams and Dale (1965) who state that for a grouping of OTU's to be considered a classification three requirements must be met (paraphrased for biological taxonomy): (1) Within every taxon containing more than one OTU there must be, for every OTU, at least another OTU with which it shares minimally one relevant character state. (2) Membership in the taxon may not itself be a relevant character. (3) Every OTU in any one taxon must differ in at least one relevant character state from every OTU in every other taxon. We must also distinguish between taxa (plural of taxon) and categories. Taxa are actual groupings observed in nature, regardless of the basis on which the grouping has been done. They are allocated to categories which are the hierarchic levels in a classificatory scheme. Thus homo sapiens, carnivores or mammals are taxa, while species, genera or families are categories.

Most classifications are internal (Williams and Dale, 1965) by which is meant that the classification is based upon criteria entirely inherent within the data that are to be classified. By contrast, there are external classificatory procedures in which certain reference taxa are employed in aiding in the classification. An example in point is the non-Linnean taxonomy of DuPraw (1964) which employs discriminant functions including both known and unknown specimens mapped in a two-dimensional space by discriminant analysis.

While classifications in psychology and the social sciences need not always be hierarchically structured, the principal purpose of biological numerical taxonomy is to group organisms into a hierarchic system of biological taxa. There are two million different species of living organisms in the world. These must be grouped if only for convenience of creating order in a chaos of names and forms, but also because a sound taxonomic system will reveal much that is useful and of interest about the evolutionary mechanisms that have given rise to the diversity of kinds of organisms existing in the world today. The principle of biological evolution is fundamental to an understanding of the nature of biological taxa and the discontinuities among them. This is reflected in the commonly accepted belief that there is just one "natural" system which, if only found, would be the obvious classification of the group under study. Traditionally this natural classification has always been an evolutionary one. Presumably the organisms constituting a taxon are related by common descent. If we could only go back in the fossil record of a natural group, we would find a common ancestor for them before encountering a common ancestor for these forms and those in another taxon of equal rank. However, it has been

emphasized in recent years that there are at least two fundamental kinds of relationships among taxonomic units, phenetic relationships which are based on overall similarity in terms of the characteristics which are measured, and cladistic relationships based on common descent as described above. Most conventionally stated taxonomic relationships contain an undefined mixture of the two (Sokal and Camin, 1965). Naturalness in a phenetic sense is understood to mean maximal overall similarity within a taxon as contrasted with substantial differences from other taxa.

Thus, in attempting to set up a natural system we have to say whether it is natural in a phenetic or a cladistic sense. Most of the work in numerical taxonomy so far has dealt with phenetic systems, taxonomies based on overall similarity which may or may not reflect closeness of evolutionary relationship. These systems are of general utility. Phenetic taxa in a natural system should be cohesive and have a high predictive value for characters other than those upon which the taxonomy has been based. This brings up the problem of character selection, which does not loom as large in sociology and psychology, because only characters of interest are chosen. Thus, if we want to classify individuals on the basis of their attitudes to drinking, we might only classify them on responses related to this variable, but would not necessarily classify them on their physiology, their attitudes to art, or their driving habits. The question is whether there are natural taxa of personality types rather than different, partially intersecting facets of the personality. In biology we wish to represent as fairly and exhaustively as we can the genetic structure of the individual populations under study, and this leads to serious problems of character selection as we shall see.

Fundamental to the establishment of any taxonomy is the decision on whether taxa are to be monothetic or polythetic. A monothetic group is defined by the possession of a unique set of features, and classification on monothetic principles is a series of successive logical divisions into ever smaller subsets sharing one or more states of a character. By contrast, a polythetic classification places together organisms that have the greatest number of shared features. No single feature is either essential to group membership or is sufficient to make an organism a member of this group.

#### Similarity coefficients

Any consideration of clustering methods must concern itself with the nature of the data to be clustered. A few of the methods extract structure directly from the original data matrix, which is a rectangular matrix whose columns are operational taxonomic units (the OTU's to be clustered) and whose rows are the characters on the basis of which the clustering proceeds. The characters are coded numerically into a number of states or as a continuous function. In the majority of cases we first compute from the data matrix a matrix of similarity coefficients, which expresses the pair-wise relationships among all the OTU's of the study. These coefficients of similarity are of three basic kinds--coefficients of association, which in some way express the measure of agreement in character states that actually exists

between any pair of OTU's as a proportion of the total amount of agreement that could exist; correlation coefficients among OTU's, based on the characters of the data matrix (this is the conventional Q-type analysis of the psychologists); and a measure of Euclidian distance between OTU's in a character-space. For purposes of this discussion I shall confine myself to a discussion of correlation and distance coefficients with which I have had most experience. Many of the association coefficients can be transformed to distances or functions thereof. An important consideration first pointed out by Williams and Dale (1965) is that while studies of the relationships among OTU's, whether measured as correlations or as distances have both been termed Q-studies following the lead of the psychometricians, there is a profound difference between these. A matrix of correlations between pairs of OTU's represents angles among OTU's in a space whose dimensions represent the OTU's. Thus, there are maximally as many dimensions as there are OTU's. On the other hand, a distance matrix among pairs of OTU's, while also a Q-study, shows distances among OTU's imbedded in a character space. That is, the dimensions of the hyperspace represent the separate characters, or, seen in the three-dimensional representations of OTU's which we have been preparing for purposes of study and analysis, these three dimensions represent linear combinations of the characters (three eigenvectors corresponding to the three largest eigenvalues of the character correlation matrix). Conversely, correlations among characters would represent angles in a character space. Distances among characters are not generally computed, but if they were, they would be imbedded in a space whose dimensions were the individuals of the study. Williams and Dale (1965) have called the character space an A-space (from attribute space) while the space whose dimensions represent the OTU's has been called an I-space (from individual space).

Several characteristics of the similarity coefficients profoundly affect the clustering methods. The similarity function should be metric, that is, it should meet the requirements of symmetry, the triangle inequality, and should be non-zero for nonidentical elements and zero for identical ones. Most coefficients proposed in numerical taxonomy have been metric. Some semimetric and asymmetric similarity coefficients have been proposed in numerical taxonomy and in some instances such as immunological similarity may be justified. However, such coefficients greatly complicate the clustering and analysis of the OTU's.

General considerations of the relations among the similarity coefficients are in order. For instance, since the taxonomic relations resulting from the cluster analysis are to be in the nature of universals, it is important that one-to-one relations between these coefficients be established, although, of course, these coefficients cannot be linear functions of each other; otherwise, there would be little point in preferring one over the other. One would at least hope that monotonicity of the similarity function is retained. In fact, however, it can be easily demonstrated that the various similarity functions so far employed in numerical taxonomy are not jointly monotonic. Decisions, therefore, have to be taken upon the choice of coefficients, based partly on the model of the type of similarity which it is desired to portray and partly on the mathematical properties of the coefficients.

Another important consideration of a classification is stressed by Williams and Dale (1965). It is not necessarily true that a given similarity function used to set up a classification at the lower hierarchic level will decrease (or increase) monotonically as we ascend the hierarchy. An example of this is the Spearman's sums of variables method which frequently leads to reversals in the value of the correlation coefficient when clusters join, as noted by Sokal and Michener (1958). Furthermore, in certain types of approaches the consequential nested hierarchies are not retained and several members of a subset at a low hierarchic level may split up to become members of different sets at a higher hierarchic level. Such relationships have been observed, among others, by Rubin (1966) in his optimal taxonomy program.

One decision that must be made is whether a similarity index is to be constructed which will indicate what is most similar to the human observer or whether such an index can measure what might be described as the intrinsic similarity between two objects based on their component parts, this latter similarity not necessarily congruent with the one apparent to the observer.

#### Clustering methods

The three main clustering methods employed in biology have been the methods described as linkage methods by Sokal and Sneath (1963). In all of these methods the criterion for joining is gradually lowered from an initial high similarity value at which all OTU's are represented by a disjoint partition (single OTU's in a subset) to low similarity values at which the classification is represented by a conjoint partition (all OTU's are in the same taxon). Single linkage described by Sneath (1957) permits a single linkage between an OTU and a cluster or between two clusters to establish a new, more inclusive cluster. While two clusters may be linked by the single linkage technique on the basis of a single bond, many of the members of the two clusters may be quite far removed from each other. To overcome this difficulty, Sneath has recommended recalculating mean similarity values both within and between groups (see Sokal and Sneath, 1963, page 181). Wirth, Estabrook and Rogers (1966) use graph theoretical techniques and representation to carry out what is essentially a single linkage method. Clustering by complete linkage requires that a given OTU or a cluster joining another cluster at a certain similarity coefficient  $S_i$  must have relations at that level or above with every member of the cluster to be joined. This yields compact and conservative clusters compared to the long, strung-out classifications of single linkage. The average linkage method calculates average similarities of clusters with prospective joiners and since its initial development by Sokal and Michener (1958) classifications based on it and on its various modifications have demonstrated higher cophenetic correlations with the original similarity coefficients than classifications based on other clustering methods.

Lockhart and Hartman (1963) have developed a technique for successively subdividing large numbers of bacterial species into groups by monothetic criteria. Their results were, in effect, similar to those obtained by



polythetic methods. The method by Camin and Sokal (1965) for clustering OTU's in preparation for cladistic analysis is another modified method. Several studies are now available comparing different methods of clustering (see Lange, Stenhouse and Offler, 1965; Williams, Lambert and Lance, 1966; and Sokal and Michener, 1967). Without discussing these studies in detail, we can summarize them by stating that different similarity coefficients as well as different clustering operations yield appreciably different phenograms from the same data. Sokal and Michener (1967) conclude that "As to clustering procedures all the different methods tried produce somewhat different results. . . .

"It is becoming clear that the procedures for clustering OTU's will need considerable scrutiny and improvement if the aim of achieving stability in classification is to be realized. Each of the methods of clustering so far tends to bias the resulting clusters in certain ways. Thus, for example, the weighted pair-group method with arithmetic averages assumes that OTU's occur in nested, dendritic clusters. It will best cluster OTU's from a similarity matrix which does in fact have such phenetic relationships and it will tend to impose dendritic relationships upon data that are not markedly dendritic. The degree to which the phenogram reflects the similarity matrix (cophenetic correlation) must indicate the degree to which the clustering method represents the underlying structure among the OTU's. It is therefore important to investigate this structure by a variety of techniques and to ascertain the nature of the phenetic constellations of OTU's in different taxonomic groups. Given an understanding of the phenetic structure of a taxonomic group, it should be possible to recommend an appropriate clustering method for it. No one clustering method is likely to serve well in every instance. To give an extreme example, members of a continuous cline clearly would not be appropriately clustered by any of the average linkage methods."

A major unresolved problem of cluster analysis in biology is the fact that few, if any, clustering methods have been devised which do not in some way bias the resulting classification. The average linkage method will attempt to give best results with hyperspheroidal clusters separated by substantial gaps, single linkage does well with strung-out data, and so forth. Moreover, these clustering methods tend to bias the resulting structures in the direction implied by the clustering procedure. It is, therefore, of considerable importance to try to establish general clustering procedures whose algorithm would vary depending on the scatter and distribution of the OTU's to be clustered. Thus, if the OTU's are in fact spheroidally clustered, the average linkage procedure might well be used. If, on the other hand, more complex shapes such as hypeserpentines, hyperdumbbells, hyperdoughnuts, or even hyperfleurs-de-lys are closer to a representation of the essential distribution of the points in hyperspace, then the clustering program should adjust itself to such patterns. Such self-adjusting programs are still not extensively developed, but it seems to me that we shall not be representing nature faithfully, nor learn much about the forces that have resulted in the phenetic patterns being observed, unless we produce programs of this sort. Rohlf (1967) has developed a clustering procedure which departs from the conventional hyperspheroid by allowing hyperellipsoid clusters, reaching out farther in some directions away from the center than in others.



To avoid the distortions necessary by the two-dimensional representation of phenograms, numerical taxonomists have recently turned increasingly to other means of representation of taxonomic relationships. Among the most popular is the three-dimensional plotting of OTU's either as models or in two-dimensional perspectives. In such plots, the dimensions usually represent the largest three eigenvectors from the character correlations and are thus linear combinations of the characters. It has been our experience that the first three factors usually extract 50-70 per cent of the overall variance. However, the cophenetic correlations (see below) between distances in the resulting three-space and the original similarity matrix are always above 0.90. Such representation leaves, of course, the actual categorization unresolved, and methods will have to be developed for handling such problems. Most recently Rohlf (1967) has developed a method for representing taxa in stereograms which give the illusion of three-dimensional projection when examined with stereoscopic glasses.

#### Some other considerations

An important question related to choice of characters is how many and which characters to choose to establish a stable natural classification. Numerical taxonomists have maintained that as the number of characters employed increases an asymptote of information is reached, and that equal increments in numbers of characters employed will provide decreasing perturbations of the taxonomy. This seems obvious from a statistical point of view if we can conceive of the characters as randomly selected from an infinite population of possible characteristics measuring similarity among a given pair of OTU's. Experiments are under way to test this hypothesis, and we are not yet in a position to render final judgment upon it. This line of argument leads, however, to a position where each sample of characters in a taxonomic study is considered equivalent to every other sample of characters, both from the point of view of importance (the assumption of equal weighting of characters in expressing similarity) as well as from the point of view of providing equivalent information about similarities. This latter point is important, because it assumes that regardless of what sets of characters we chose, be these external or internal morphological characters as well as biochemical or physiological characters, we should be able to obtain identical taxonomies. Investigations of this hypothesis of non-specificity by Rohlf (1963) and Michener and Sokal (1966) have shown that different sets of characters will yield similar but not identical classifications, measures of the replicability of the classification yielding cophenetic correlations between 0.42 and 0.85.

Results from these studies as well as from another study in which independent investigators reclassified identical sets of objects lead to the recognition of what Rohlf has called the uncertainty principle in taxonomy. This simply states that it is impossible to reclassify by conventional or numerical means the same set of organisms and obtain comparable results beyond a certain degree of replicability. The resemblance among successive classifications may be very great (cophenetic correlations on the order of 0.85). On the other hand, the uncertainty may be considerably greater. Our experience in this field has not yet been sufficient to indicate between which bounds this uncertainty may lie.

On what criterion can a classification be judged? In the early days of numerical taxonomy, the success of a numerical classification was generally judged by the similarity of the outcome to those classifications established by conventional means. As the subject developed, there seemed no inherent reason why the traditional, somewhat intuitive, classifications should be considered as the final arbiter, and attempts were made to develop internally sufficient criteria for the goodness of a classification. Two main approaches have been followed. Sokal and Rohlf (1962) have used the method of cophenetic correlation which consists of correlating the original similarity matrix with so-called cophenetic values which are the values of similarity implied by the structure of a given classificatory phenogram. Phenograms are two-dimensional representations of taxonomic structure in terms of trees with the axis parallel to the stem of the tree representing phenetic similarity. Because phenograms collapse multidimensional relationships into two dimensions, there is appreciable distortion of the original relationships as shown in the similarity matrix. The goodness of a classification can now be measured as magnitude of the correlation between a phenogram and the original similarity matrix. It is, of course, desired that the phenogram represent as much as possible the phenetic similarity as shown in the similarity matrix. Of two taxonomic representations based on the same similarity matrix, that with the higher cophenetic correlation is to be preferred. A method recently developed by Rohlf (1967) permits the moving of some of the branches by a trial-and-error basis into positions yielding higher cophenetic correlations. However, this procedure is not yet practical for very large matrices, except on exceedingly fast computers.

Rubin (1966) has approached the subject from the general point of view of establishing a stability function for a given classification, which is to be a measure of the homogeneity within groups and the inhomogeneity among groups at a given hierarchic level. Once such a function can be defined, one obviously wishes to maximize it, that is, one wishes to arrange the OTU's within a classification in such a way that the function becomes maximized. Since any given classificatory procedure will not result in maximization of the function, rearrangement of the OTU's among the classes to yield an improved classification can be attempted by a variety of algorithms. Rubin's hill-climbing algorithm proceeds to follow up improvements of his stability criterion. In fact, once such a criterion for stability or goodness of a classification is accepted, then almost any randomly chosen classification of objects can be successively improved by a series of iterative steps yielding successively higher criteria.

Of special interest are some types of self-adjusting clustering methods which have been described in the literature. These include conceptually simple, but computationally complex methods such as curves derived from scattered points which represent the essential trends of these points (Sneath, 1966). Other techniques seek by some method of cluster analysis to classify a series of OTU's and subsequently, using each OTU as an improvement of the previous classification, allocating it to previously established classes unless it would seriously disagree with the established classificatory scheme (Ornstein, 1965). This scheme is essentially a "learning" classification program improving its performance for a given set of data after having initially classified a certain number.

## Literature Cited

- Ball, G. H. 1965. Data analysis in the social sciences: What about the details? Proc. Fall Joint Comp. Conf. 533-559.
- Camin, J. H. and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19:311-326.
- Dupraw, E. J. 1964. Non-Linnean taxonomy. *Nature* 202:849-852.
- Lange, R. T., N. S. Stenhouse and C. E. Offler. 1965. Experimental appraisal of certain procedures for the classification of data. *Australian Journal of Biological Sciences* 18:1189-1205.
- Lockhart, W. R. and P. A. Hartman. 1963. Formation of monothetic groups in quantitative bacterial taxonomy. *Journal of Bacteriology* 85:68-77.
- Michener, C. D. and R. R. Sokal. 1966. Two tests of the hypothesis of nonspecificity in the Hoplitis complex (Hymenoptera: Megachilidae). *Annals of the Entomological Society of America* 59:1211-1217.
- Ornstein, L. 1965. Computer learning and the scientific method: A proposed solution to the information theoretical problem of meaning. *Journal of the Mount Sinai Hospital* 32:437-491.
- Rohlf, F. J. 1963. Congruence of larval and adult classifications in Aedes (Diptera: Culicidae). *Systematic Zoology* 12:97-117.
- Rohlf, F. J. 1967. Manuscript in preparation.
- Rubin, J. 1966. An approach to organizing data into homogeneous groups. *Systematic Zoology* 15:169-183.
- Sneath, P. H. A. 1957. The application of computers to taxonomy. *Journal of General Microbiology* 17:201-226.
- Sneath, P. H. A. 1966. A method for curve seeking from scattered points. *The Computer Journal* 8:383-390.
- Sokal, R. R. and J. H. Camin. 1965. The two taxonomies: Areas of agreement and conflict. *Systematic Zoology* 14:176-195.
- Sokal, R. R. and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38:1409-1438.
- Sokal, R. R. and C. D. Michener. 1967. The effects of different numerical techniques on the phenetic classification of bees of the Hoplitis complex (Megachilidae). *Proceedings of the Linnean Society, London* (in press).

- Sokal, R. R. and F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon* 11:33-40.
- Sokal, R. R. and P. H. A. Sneath. 1963. *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco and London. 359 p.
- Williams, W. T. and M. B. Dale. 1965. Fundamental problems in numerical taxonomy. *Advances in Botanical Research* 2:35-68.
- Williams, W. T., J. M. Lambert and G. N. Lance. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. *The Journal of Ecology* 54:427-445.
- Wirth, M., G. F. Estabrook and D. J. Rogers. 1966. A graph theory model for systematic biology, with an example for the *Oncidiinae* (Orchidaceae). *Systematic Zoology* 15:59-69.

# A Mutual Development of Theory and Method in Objective Analysis of Personality Structure

Louis L. McQuitty  
Michigan State University

In September of 1965 I summarized my approaches up to that time under the title "A Mutual Development of Some Typological Theories and Pattern-Analytic Methods"(McQuitty, 1967).

I wish now to review more recent developments in my approaches. These are a continuation of the earlier approaches and are introduced here by summarizing my position in both theory and methods as of the close of my earlier review paper.

## A Brief Review

### General

My general approach is to develop methods of analysis out of theories of personality structure. Applications of methods to data serve as hypotheses for testing theory. They lead to the revision of theory and the development of new methods. Through this approach, I attempt to develop both better theory and improved methods.

My theoretical position as of September 1965 was as follows:

- "(1) Every person is an 'imperfect' type as distinct from a 'pure' type; only 'imperfect' types exist in reality, and 'pure' types exist only in theory.
- (2) There are fewer 'pure' types than 'imperfect' types; each 'pure' type is represented in reality by two or more 'imperfect' types.
- (3) The characteristics of 'pure' types are approached but never quite realized by classifying 'imperfect' types into internally-consistent categories, and determining their common characteristics. The validity of representation of a 'pure' type generally increases as the number of 'imperfect' types representing it increases.
- (4) 'Hierarchical' types include all of the types realized in classifying 'imperfect' types into larger and larger, internally-consistent categories; they are the types intermediate between those of reality and theory, 'imperfect' and 'pure' " (McQuitty, 1966a).

A category of persons is said to exemplify a statistical type if everyone in the category is more like every other person in the category than he is like any person in any other category.

In converting the theory of types to a method of analysis, persons are described by patterns of characteristics which they possess. An index is computed showing the degree of relationship of every person to every other person in terms of common characteristics and the results are assembled in a matrix. The matrix reports an index of similarity of every person to every other person.

In accordance with the definition of types, the matrix is searched for internally-consistent submatrices. A submatrix of two persons is internally consistent if Individual  $i$  is most like Individual  $j$  and  $j$  is in turn most like  $i$ . Internally-consistent submatrices of higher order are defined analogously.

Internally-consistent submatrices of any size can be isolated by the methods of Reciprocal Pairs or Rank Order Typal Analysis, as described elsewhere (McQuitty, 1964 and 1966a).

Each internally-consistent submatrix defines a hierarchical type. Each hierarchical type has the characteristics which are common to its members. Each hierarchical type is assumed to be a better representative of a pure type than is any one of the imperfect types with which it is compared.

The imperfect types of the internally-consistent submatrix are replaced in the original matrix by the hierarchical type, and the analysis proceeds in this fashion until all persons are classified into one of two major hierarchical types as shown in Figure 1.

Insert Figure 1 about here

#### An Illustration

An example, using Hierarchical Analysis by Reciprocal Pairs, will help clarify the general approach. The method was applied to a matrix of agreement scores between industrial companies which had been analyzed many times in terms of other pattern analytic methods. The agreement scores for these companies are shown in Table 1.

Insert Table 1 about here

Variables A and B represent two construction companies, C and D trucking companies, E and F grain processing and metal products respectively, and G and H garment companies with female employees only; the other six companies employed male employees only. The companies were assessed in terms of 32 variables. Each variable was dichotomized at the median and two companies agreed on a variable if they were both either above or below the median, but not if one was above and the other below the median. The agreement score (Zubin, 1938) is the number of items on which the two companies agree.

The reciprocal pairs of Table 1 are underlined. They are for Pairs AB, CD, EF, and GH. Company A, for example, has Company B most like it, and Company B in turn has Company A most like it, thus fulfilling the requirements of reciprocity as used here.

Companies A and B have in common 29 of the 32 characteristics on which they were assessed. These two companies are collapsed into a single hierarchical type AB and are characterized by their 29 common characteristics. In a similar fashion members of the other three reciprocal pairs are collapsed into three Hierarchical Types, CD, EF, and GH, described by 26, 21, and 24, common characteristics respectively.

The agreement score of every hierarchical type with every other hierarchical type is computed and the results are reported in Table 2. Table 2 is analyzed in

Insert Table 2 about here

the same fashion as Table 1. Results of the analysis of the two tables are shown in Figure 1.

### Purifying the Data

Hierarchical Classification by Reciprocal Pairs attempts to purify the data in relation to types as the analysis proceeds (McQuitty, 1966a). Lower level types are assumed to be more imperfect than higher level types. Consequently, when any two imperfect types such as E and F of Figure 1 are combined into a single higher type, EF, this latter type is assumed to possess only the characteristics which the two imperfect types, E and F, have in common.

The above assumption was applied to the agreement scores of Table 1 and produced the classification reported in Figure 1.

That Hierarchical Classification by Reciprocal Pairs does in fact sometimes purify the data can be illustrated by comparing the results from it with analogous results from Rank Order Typal Analysis.

Rank Order Typal Analysis makes no effort to purify data as it proceeds.

The first step in Rank Order Typal Analysis is to convert the data of Table 1, for example, into ranks within columns, where the highest rank of each column is assumed to be the entry  $r_{ii}$  for the reliability of each person with himself.

The ranks within columns of the data of Table 1 are shown in Table 3. This

Insert Table 3 about here

latter table shows in Column C, for example, that C is assumed to be most like C and is therefore assigned a rank of 1. The other ranks are assigned in terms of the relative size of the agreement scores in Table 1. Company D has the largest agreement score with C (except for C with itself) and is therefore assigned a rank of 2. Other ranks are assigned in an analogous fashion.

A Rank Order Analysis of Table 3, as reported earlier (McQuitty, 1963) produces the results shown in Figure 2.

Insert Figure 2 about here

Inspection of Table 3 shows that Company E is most like F and in turn has F most like E. The two companies form a type as shown in Figure 2. The analogous result is obtained for Companies G and H.

The classifications differ, however, from those obtained with Hierarchical Classification by Reciprocal Pairs. Companies E, F, G, and H do not form a type in Rank Order Typal Analysis but they do form a type in Hierarchical Classification by Reciprocal Pairs.

The difference in the two approaches is emphasized by comparing Table 2 with Table 3.

Using Rank Order Typal Analysis, Table 3 shows that Companies E, F, G, and H do not form an internally-consistent category in being like one another. If they did, there would be no rank larger than four, the number of cases in the submatrix EFGH by EFGH.

On the other hand, when the four companies E, F, G, and H are first classified into two hierarchical types, EF and GH, and are thereby purified in the method of

Hierarchical Classification by Reciprocal Pairs, they then yield an internally-consistent pair of hierarchical types as shown in Table 2; this justifies their classification into a hierarchical type, EFGH.

In this example, Hierarchical Classification purified the data as it proceeded in the analysis.

#### Some Limitations of Hierarchical Classification by Reciprocal Pairs

Although Hierarchical Classification has many advantages as outlined elsewhere (McQuitty, 1964, 1965, 1966a, 1966b, 1967) it has certain limitations.

The initial classification begins at the bottom of the hierarchical system and depends primarily on only a few of all of the indices of association in a matrix. Mistakes might occur early in the analysis as a result of using only a few relatively unreliable indices and might have serious consequences for the subsequent classifications.

#### Two Approaches toward a Solution

There are two possible attacks on these problems. One approach is to attempt to increase the reliability and validity of the few indices on which the classification decisions depend.

Another attack on the problem is to attempt to develop a method which starts at the top of the hierarchical system, uses all indices, and builds downward. Such an approach might divide the original matrix into two submatrices and then continue by dividing the successive submatrices until a structure such as represented in Figure 1 is built from the top down.

#### A Joint Solution

##### General Description

In attempting first to solve only the first problem, viz., to increase the validity of the few indices on which the classifications depend, I discovered an approach which solves both this problem and the one of using all indices in each decision. The new method divides the large matrix into submatrices and then divides successively each submatrix using in each case all of the indices of the matrix or submatrix on which the operations are performed. Each time before making a division, the method takes steps designed to increase the reliability and validity of all indices.

##### Detailed Description

The method is now described in more detail in the order in which it was developed.



Increasing the Validity of Indices. In attempting to increase the reliability and validity of the  $r_{ij}$  in an  $N$  by  $N$  matrix of indices between people, the correlation was computed between corresponding entries of the columns  $i$  and  $j$ . This approach gave an index of the extent to which  $i$  and  $j$  varied jointly in relation to the other  $N-2$  variables of the matrix. In other words, the relationship between  $i$  and  $j$  is estimated by computing the extent to which they are jointly like  $N-2$  other variables. The new index is called an intercolumnar correlation, and is designated  $I$ .

In testing the validity of the intercolumnar correlations, as compared with agreement scores, intercolumnar scores were computed for the agreement scores of Table 1.

The Pearsonian Coefficient was used in computing the intercolumnar correlations between columns of agreement scores.

In testing the validity of the intercolumnar correlations, they were pattern analyzed and compared with previous pattern analyses of agreement scores. Rank Order Typal Analysis of intercolumnar coefficients gave the same results as Hierarchical Analysis by Reciprocal Pairs.

The computation of intercolumnar correlations from agreement scores purified the data in a fashion somewhat similar to Hierarchical Classification by Reciprocal Pairs; the two approaches produced identical classifications.

Based on both other pattern analytic analyses of the data and known characteristics of the companies, the Rank Order Typal Analysis of the intercolumnar coefficients produced a more valid picture of the structure than did the Rank Order Typal Analysis of the agreement scores.

Seeking the Nature of the Improvement. To seek the nature of the error being corrected by (a) intercolumnar correlations and (b) Hierarchical Classification by Reciprocal Pairs seemed worthwhile.

As a first step in this direction, Pearsonian Coefficients between the several companies were computed on the basis of the original 32 scales for the eight companies, to yield the matrix shown in Table 4.

Insert Table 4 about here

This table was converted to ranks within columns and produced the same results exactly as did the ranks within columns of the intercolumnar correlation of agreement scores. The ranks for both approaches are shown in Table 5.

Insert Table 5 about here

An inspection of this table shows that a Rank Order Typal Analysis of it will yield the same classification as obtained by (a) Hierarchical Classification by Reciprocal Pairs when applied to the agreement scores and (b) Rank Order Typal Analysis of Intercolumnar Correlations of Agreement Scores (Figure 1).

In summary, both (a) the computation of intercolumnar correlations, and (b) Hierarchical Classification by Reciprocal Pairs corrected errors introduced when the data were dichotomized and agreement scores computed to represent the data.

The above results support the hypothesis that intercolumnar correlations of agreement scores between people are more valid for the isolation of types than are the agreement scores themselves.

#### A Statistical Method Generated by a Hypothesis

##### The Hypothesis

The above hypothesis was used to generate another hypothesis. If the first intercolumnar correlations of original indices of a matrix enhance the emergence of types then possibly the next and subsequent computations of intercolumnar correlations would facilitate still farther the appearance of types if they are present but hidden in the original indices. It is therefore hypothesized for further study that iteration of intercolumnar correlations generates the emergence of types in a matrix of interassociations between people if types are present but hidden in the original matrix.

##### Testing the Hypothesis

First Test. The same set of data was used in testing this hypothesis. The standing of the eight companies on the 32 scales was used in lieu of dichotomized

data. The Pearsonian Coefficient of correlation for every company with every other company was computed. The results are reported in Table 4.

The intercolumnar Pearsonian Coefficient of correlation was then computed for every column with every other column of Table 4, to yield the first intercolumnar matrix. The process was repeated on the first and subsequent intercolumnar matrices until in the fifth matrix all entries became either plus one or minus one as shown in Table 6.

Insert Table 6 about here

Table 6 reflects two types, ABCD and EFGH, each defined by a submatrix in which all entries are plus one.

The above procedure was applied to variables of each submatrix, using in each case the original entries of correlation reported in Table 4. Again each submatrix was divided into two smaller submatrices of plus one entries. The process isolated Types AB and CD in the third intercolumnar table of Submatrix ABCD and Types EF and GH in the fourth intercolumnar table of Submatrix EFGH.

The original correlations and the three submatrices of intercolumnar correlations for Variables A, B, C, and D are shown in Table 7.

Insert Table 7 about here

The above Iterative Intercolumnar Correlational Analysis produced the same types as did both (a) the Hierarchical Classification by Reciprocal Pairs and (b) the Rank Order Typal Analysis of the first intercolumnar correlations from agreement scores. The results for the three analyses are shown in Figure 1.

Second Test. As a more crucial test of the ability of Iterative Intercolumnar Correlational Analysis to yield types if operative in the data, the method was applied to a set of data which earlier proved relatively resistant to pattern-analytic methods. The data are particularly difficult to pattern-analyze because they include many ties in crucial agreement scores.

"The data were generated by requesting a subject to react to the pictures of 20 art objects, by using adjectives which might describe them (40 adjectives were used). For each art object the subject went through the entire list of adjectives before proceeding to the next object. The subject responded by saying, in effect, that the adjective is descriptive of the object; that it is not descriptive; or that she could not decide whether or not it is descriptive. If the subject's initial response to a picture was positive, she then endorsed one of three alternative answers: (1) 'I like the characteristic described by this adjective,' (2) 'I do not like it,' (3) 'I can't decide whether or not I like it.' ...

"An agreement score (Zubin, 1938) was computed for every object with every other object. Suppose that there were six adjectives and two objects and that the subject reported the following reactions:

Adjectives	1	2	3	4	5	6
Object A	Yes, Like	Yes, Dislike	Yes, ?	No	No	Yes, Dislike
Object B	Yes, Like	?	Yes, ?	Yes, Like	No	Yes, Dislike

The agreement score between A and B for these six adjectives would be four, the agreement being on Items 1, 3, 5, and 6 only.

"Similar computations across all 40 adjectives and among all 20 objects yielded the 20 x 20 matrix of agreement scores shown in" Table 8 (McQuitty, Price, and Clark, 1967).

Insert Table 8 about here

Table 9 reports the first matrix of intercolumnar correlations and Table 10 reports the fifth matrix of intercolumnar correlations, viz., the first table in which all entries were either plus or minus one, to yield Types CFIMPANQEKs and JTRGBOHDL.

Insert Tables 9 and 10 about here

Figure 3 shows the results from the complete analysis of the data by Iterative Intercolumnar Correlational Analysis. These results show that the current method can

Insert Figure 3 about here

analyze with ease one set of data which has proven difficult for most methods because the data involves tied values crucial for several required decisions.

General Evaluation. Every classificatory decision in the iterative method is based on all of the indices of the matrix being analyzed as compared with primarily only one index in the reciprocal pairs method. It is, therefore, hypothesized to be both more reliable and valid than the latter method. These points need further study.

#### A Mathematical Proof

There is a more sophisticated approach to substantiating the hypothesis that Iterative Intercolumnar Correlational Analysis will isolate types if operative in the data, viz., to prove it mathematically.

The Generation of Plus One Intercolumnar Correlations. In the development of the proof, a type is defined as a category of people of such a nature that everyone in the category has a group of common characteristics, and anyone not in the category does not possess all of these characteristics.

Assume now that we have a matrix of interassociations between people based on test item responses which assess typal membership with validity better than chance. Assume also that any variance in the responses to the test which is not attributable to typal membership is governed by chance alone.

Let:

- 1)  $i$  and  $j$  be any two individuals of the same type.
- 2)  $x_1, x_2, x_3 \dots x_N$  be any  $N$  individuals with no one of them specified in any way as to typal membership.

The coefficients of correlation between these variables are indicated in Table 11.

Insert Table 11 about here

Let:

$N$  be infinitely large, so large that chance variation is ignored.

$$3) \bar{r}_{x_1 i} = \bar{r}_{x_1 j}; \bar{r}_{x_2 i} = \bar{r}_{x_2 j}; \bar{r}_{x_3 i} = \bar{r}_{x_3 j} \dots \bar{r}_{x_N i} = \bar{r}_{x_N j}$$

4) Let  $I_{ij}^{1 \rightarrow N}$  = the intercolumnar correlation between  $i$  and  $j$ , i.e., the correlation between corresponding entries of columns  $i$  and  $j$  of Table II.

If the entries for Columns  $i$  and  $j$  of Table II were known, the intercolumnar coefficient could be computed by substituting the entries of Columns  $i$  and  $j$  in a regular formula for computing the Pearsonian  $r$ .

Analogously, the symbols of Columns  $i$  and  $j$  can be substituted in a raw score formula for computing  $r$ . This new formula is then the intercolumnar coefficient  $I_{ij}^{1 \rightarrow N}$ . This new formula can be simplified by substituting either the  $r_{x_1 i}$  for the corresponding  $r_{x_1 j}$  or the  $r_{x_1 j}$  for the corresponding  $r_{x_1 i}$  (from Equation 3).

5) In the first case  $I_{ij}^{1 \rightarrow N} = 1$ , except when  $\bar{r}_{x_1 i} = \bar{r}_{x_2 i} = \bar{r}_{x_3 i} \dots = \bar{r}_{x_N i}$

and in the latter case except for equality among all  $\bar{r}_{x_1 j}$ .

The above conditions would occur if and only if either all  $x$ 's belong to the same types or all  $x$ 's had nothing in common with either  $i$  or  $j$ . The proof is developed in detail elsewhere (McQuitty, \*\*).

The proof means that Iterative Intercolumnar Correlational Analysis can isolate the types reflected in a matrix of interassociation between people provided the assumptions out of which the proof developed are satisfied. Whether or not they are satisfied by the data is indicated by applying the method to the data. If the types are isolated, a search for the common characteristics of the members of each type will determine whether or not the assumptions have been satisfied. A cross validation study is required to investigate the stability of the types.

The Generation of  $N$  Intercolumnar Correlations. Another proof must be added to the above developments if the isolated types are to be easily recognized. The additional proof must show that iteration of indices for variables not in the same type will not move them to plus one as a limit.

Let:

Individuals  $w$  and  $v$  be any two individuals of "opposite" types. Two types are "opposite" if they have no common characteristics; each type has its own characteristics and lacks all of the characteristics of the other type. In data where absent characteristics are relatively meaningless from the point of view of the theory being applied, a more appropriate term would be independent rather than "opposite" types.

In this case:

$$6) \bar{r}_{x_1 u} = -\bar{r}_{x_1 v}; \bar{r}_{x_2 u} = -\bar{r}_{x_2 v}; \bar{r}_{x_3 u} = -\bar{r}_{x_3 v} \dots \bar{r}_{x_N u} = -\bar{r}_{x_N v}.$$

As before, by substituting in a raw score formula for the computation of the Pearsonian Coefficient and simplifying, the result shows that the intercolumnar correlation for two individuals of "opposite" types (over  $N$  other individuals not all of the same type) equals minus one.

The above proofs show that two individuals of the same type will yield a plus one and two individuals of "opposite" types will yield a minus one intercolumnar correlation when computed over any  $N$  individuals of more than one type.

The Generation of Intercolumnar  $r$ 's  $< +1 > -1$ . Any two individuals, not of the same type and not of "opposite" types will yield an intercolumnar correlation of less than plus one and greater than minus one when computed over any  $N$  individuals. This is because data of this kind cannot satisfy either Equation 3 or 6; the first of these equations must be satisfied if the intercolumnar correlation is to be plus one and the second equation must be satisfied if it is to be minus one.

The above developments show that the intercolumnar correlation between any two individuals is less than plus one and greater than minus one, if computed over other individuals all of the same type, and also when computed over individuals of different types, except when the two individuals are either of the same or opposite types. When the two individuals are of the same type, the intercolumnar correlation is plus one, and it is minus one when they are of opposite types.

#### The Reverse Proofs

The reference article (McQuitty, (\*\*\*)) shows also that the proofs can be reversed to show:

1. If the intercolumnar correlation between two individuals  $i$  and  $j$  is plus one, then they belong to the same type.
2. If the intercolumnar correlation between two individuals  $x$  and  $y$  is minus one, then they belong to "opposite" types, where "opposite" is defined to mean that each type has characteristics of its own and each lacks all the characteristics possessed by the other; there is no overlap of the typical characteristics.
3. If the intercolumnar correlation between two individuals,  $m$  and  $n$  is less than plus one and greater than minus one, then  $m$  and  $n$  have not been proven to be members of either a single type or of "opposite" types.

The above developments show that iterative intercolumnar analysis can be used to isolate types, as already illustrated in this paper with real data.

Further elaborations of Intercolumnar Correlational Analysis are reported elsewhere (McQuitty, \*\*\*, and \*\*\*\* ).

#### Suggested Advantages of the Method

Suggested advantages of the method are: (1) it is rapid, simple to program for a computer, and can be applied to large sets of data when electronic computers are used; (2) the method uses all available, pertinent data; (3) the analysis proceeds by first dividing a matrix of associations between people (or other objects) into major submatrices and then redividing these and subsequent submatrices into smaller and smaller submatrices until all types are defined by submatrices;

(4) the method implicitly hypothesizes internally consistent types in data and either substantiates or fails to substantiate the hypothesis; (5) the raw data is required to be internally consistent within only broad chance limits; (6) the method yields a simple structure (if the hypothesis of internally consistent types is substantiated) where simple structure is defined to mean correlations of plus one between all members of every type, and less than plus one down to and including minus one between types.

#### Limitations of the Method

Even Iterative Intercolumnar Correlational Analysis, with all of its suggested advantages, does not solve all of the problems in the isolation of types.

One particularly difficult problem is the fact that indices of association between people vary with the test items used in assessing them. Consequently, the types into which people classify vary with the test items used in assessing the people.

#### The Problem of the Single Response by the Single Subject

In an effort to solve this problem, I have addressed myself first to a simpler and more fundamental problem, the problem of interpreting a single response by a single subject.

"One of the problems of interpreting a response to an item of a test is that it can be assigned various meanings depending on both who gives it and the other responses (to other items) with which it occurs.

"A single response with variable meanings can be found to have stability in psychological space if it can be assigned to a combination of responses which has stability. The response can, however, still have a kind of variability, for it might be assigned to several combinations of responses, and each of them might have stability.

"In other words, I attempt to account for the variability of meaning of a response by assigning it to several combinations of responses, each of which has stability in theoretical psychological space.

"The term psychological space is used to emphasize the possibility that identical responses (objectively) might prove to have various psychological meanings.

#### Inter- and Intra-Individual Differential Validity

"Elsewhere, I have used the term differential validity to refer to the possibility that a response might assess different attributes in different persons (McQuitty, 1959).

"Differential validity is involved (as illustrated in Figure 4) when a given Response i is endorsed along with Responses j, k, and l by one category of subjects,

Insert Figure 4: about here

A, to indicate Type X and the same objective Response, i, is endorsed along with Responses r, s, and l by another category of subjects, B, to indicate Type Y. In



the first case, endorsement of Response  $i$  indicates Type X and in the latter case, Response  $i$  indicates Type Y.

"The present paper refines further the concept of differential validity by introducing two forms of it, inter- and intra-individual.

"Inter-differential validity is now used to mean what we intended originally by differential validity, as summarized above.

"We recognize now the possibility that a response may be applied in a typological theory to assist in assessing various attributes in the same individual, depending upon the other combination of responses with which it is interpreted.

"In introducing intra-differential validity, let us suppose (as illustrated in Figure 4) that a third category of subjects, C, is formed by combining the members of each Categories A and B; they are portrayed by the common Responses  $i$  and  $l$ ," and they indicate Type Z.

"A type (such as X, Y, or Z) is defined by all of the common ways in which the members of the type behave. Consequently, each Type X, Y, and Z, would differ from each of the other two types.

"From a typological point of view (which classifies people in terms of combinations of responses), Response  $i$  with Responses  $j$ ,  $k$ , and  $l$ , indicates Type X;  $i$  with  $r$ ,  $s$ , and  $l$  indicates Type Y, and  $i$  with  $l$  only indicates Type Z. Response  $i$  would have various meanings within the same individual depending on the combination of other responses with which it is interpreted. This is what we mean by intra-differential validity, a single response assessing various attributes in the same individual depending on the other responses with which it is interpreted."

#### The Problem of a Set of Responses by a Single Individual

"A set of responses by an individual to the items of a test invites scientific explanation and understanding in the same fashion as does the single response to a single item. A set of responses has additional attractive characteristics; (1) the set can be used to assist in assigning meaning to individual responses, and (2) the set of responses can possibly be assigned to sub-sets which have relatively stable meanings.

"The problem is to devise methods for isolating all of the major and meaningful sub-sets in which the responses of an individual to a theoretically meaningful test can be assigned.

"In summary, a response may possibly have different assessment indicants for each major and meaningful combination of responses to which it can be assigned.

#### Two Solutions

"Two kinds of pattern analyses (or factor analyses) of the responses by a single individual can be recognized: (a) individual based, and (b) group based.

#### Individual Based

"In the first instance, the investigator gathers data in such a fashion that he can compute an index of the interrelation of every response by a subject to every

other response by that subject, without the use of a reference group.

"The following test items from a study now in progress fulfill the above requirement and illustrate the kind of data required for one kind of a statistical analysis of a single individual, viz., an individually based approach:

<u>Question</u>	<u>Answer Alternatives</u>		
The word angel suggests love	yes	no	?
The word angel suggests hate	yes	no	?
The word devil suggests love	yes	no	?
The word devil suggests hate	yes	no	?

"Other questions of the same kind follow with only the emotion (love, hate, etc.) changing as we move from one question to another. With this kind of an approach it is possible to compute an index of the extent to which an individual responded to angel in the same way as he did to devil.

"Using many words (in the same fashion as illustrated above for angel and devil) one can compute a matrix of interassociations between selected concepts over selected emotions. The matrix can be pattern analyzed (or factor analyzed) by any one of the many available methods. Examples of the above approach are found in studies by Schubert (1965) and McQuitty, Price, and Clark (1967)." (McQuitty, \* ).

The above described test and its method of analysis both grew out of a theory of the nature of both mental illness and mental health. The approach is described elsewhere (McQuitty, Abeles, and Clark, study in progress).

#### Group Based

"Assumptions. A series of assumptions suggests and justifies a solution to the problem of isolating the major patterns of responses of a single individual to the items of a test, as these are reflected in the responses of a group of subjects.

"We assume that every individual is an imperfect representative of one or more pure types. If two or more imperfect representatives of the same pure type are considered jointly, they give a better picture of the pure type than any one of them separately.

"If an individual is representative of n pure types, then in order to give a comprehensive, typological picture of the individual, he must be treated jointly with at least one other representative of each of these pure types.

"If a set of responses by an individual is to be understood from a group-based typological point of view, then we require a classification of the individual with one or more members of each type represented in the set of responses by the single individual. The classification is more helpful if it specifies the responses which classify the individual into each of the types he represents.

"Method. The goals implied above can be easily realized by any one of many pattern-analytic methods (McQuitty, 1967), provided only that a simple operation be introduced at the beginning of the analysis.

"Suppose that we wish to study the pattern of responses of Individual A to the items of Test X. One approach is to administer the test to 100 other individuals,

representing a universe which is meaningful in an effort to understand Individual A.

"The novel operation required by the approach of this paper is as follows: Pair the pattern of responses of Individual A with those of each of the 100 other individuals to yield Pairs: A1, A2, A3 --- A100. Specify new patterns for each of the 100 pairs, by taking the common responses of each pair. For example, if the responses by Individual A and Individual 20 were as shown in Table 12<sup>1</sup>, then Pattern A-20 (for Individuals A and 20 treated jointly) would be + on Items 1, - on 2, - on 5, + on 6, + on 7, and - on 9, with Items 3, 4, and 8 omitted because Individuals A and 20 disagree on each of these latter three items.

Insert Table 12 about here

#### Illustration

"In order to illustrate the isolation of major response patterns for a single individual, we have chosen to analyze the course selections in psychology by a single individual in relation to the course selections in psychology by the 135 other majors in that discipline, who graduated at Michigan State University during the academic years 1961-62 and 1962-63.

"During his four years of college, the one subject chosen for analysis (Code #83) registered in and obtained grades in" (McQuitty, \* ) 17 psychology courses on the quarter system. In addition to the above courses, the 135 other students majoring in psychology completed and received grades in one or more of 23 other quarter-length courses.

"The purpose is to classify the course selections by Subject A into their major, meaningful patterns, using the course selections by the other 135 students of the study as the source of information which enables us to accomplish the task.

"Individual #83 is first paired with each of the 135 other individuals of the study to yield Pairs 83-1, 83-2, 83-3 --- 83-136 (omitting 83-83). Then the courses selected jointly by the members of each pair are determined to yield patterns 83-1, 83-2, 83-3 --- 83-136 (omitting 83-83).

"An agreement score is computed between every pattern with every other pattern. For example, if Patterns 83-1 and 83-2 include the course selections shown in Table 13, then their agreement score for these five courses would be 3, the number

Insert Table 13 about here

of courses which occur in each of the two patterns (specifically courses 2, 8 and 16).

"Using the agreement scores, a matrix was prepared to show the agreement score of every pattern with every other pattern.

"Single Hierarchical Analysis by Reciprocal Pairs was applied to pattern analyze the Matrix (McQuitty, 1966a). Five individuals were chosen at random from the above group of 136 subjects, and the results from each of them were analyzed separately by the above methods.

### Results

"Four of the five individuals analyzed yielded four and only four clusters. The other individual (Code #83) produced five clusters. We elected to describe results from this individual in detail because he shows less interrelation with the other 135 individuals (in terms of the number of clusters): if his results are meaningful, then those of the others are likely also to be meaningful.

"The results for Individual 83 are shown in Figures 5 - 9 and Tables 14 - 18. Figure 5 portrays Clusters 1 and 2. Figures 6 and 7 report Clusters 3 and 4 respectively, and Figures 8 and 9 each report approximately one-half of Cluster 5.

Insert Figures 5 - 9 and Tables 14 - 18 about here

"In the first step of the analysis of the matrix, Individual 60 joined Individual 130 as shown in Cluster 2, Figure 5. Individual 5 joined Individual 39 (Cluster 5, Figure 8) and Individual 21 joined Individual 125 (Cluster 5, Figure 9). The members of each pair agreed in having selected 11 courses in common, but not common from pair to pair.

"Since only course selections used by Individual 83 were included in the analysis, Individual 83 is included in each of the above pairs and in every other combination of individuals as shown by the intersection of lines throughout the figures.

"Table 14, for example, lists certain courses (titles and code numbers) involved in Cluster 1. The body of the table shows courses which are common to each major intersection point; Courses 2, 5, 8, and 28 were selected by Subjects 35, 36, 129, and 48 to yield Intersection Point A, as shown in Figure 5, and Table 14. Intersection points involving more than five courses (but relatively few students) were omitted. Courses are reported in the tables for all of the intersection points which are labeled by capital letters in the figures. The other tables are interpreted in an analogous fashion.

"In addition to intersection points reflecting patterns of course selection, whenever a series of points is joined by a straight line parallel to the base line, all subjects of the points thus connected selected a single pattern of courses.

"One individual, Code #68, failed to appear in the analysis. He took only one course at Michigan State University in common with Individual 83. He transferred to MSU after having received credit in psychology courses elsewhere; the records do not show the specific MSU courses for which he received credit upon transfer. He was not an appropriate member of a universe in terms of which to study Individual 83. We left such individuals in the study because there were only a few of them and we wished to indicate that they would not have a major effect.

### Interpretation

"Clusters 2, 3, and 4 appear to be more meaningful than Clusters 1 and 5. Cluster 2 portrays a central interest in personality-clinical as related to psychology in business. Cluster 3 reflects an interest in individual differences in personality. Cluster 4 seems to be concerned primarily with understanding the

dynamics of the developing individual. Cluster 5 seems to be concerned primarily with the understanding of personality from a more general point of view as contrasted with a more dynamic, developmental point of view in Cluster 4.

"Cluster 1 appears to encompass personality from an experimental point of view; Course 8, Learning and Motivation, is an experimental course.

"Individual 83 appears to center his interest in personality or courses related thereto. This point is further substantiated by referring back to the courses which the subject did not select; they are less concerned with personality than are the courses which he selected.

"We conclude that the clusters are in general meaningful, and that the method has possible values as indicated further by the following development of the method.

#### Differential Pattern Analysis

"The method can be expanded to do for patterns what item analysis does for items. Item analysis selects the items most highly related to a criterion. Analogously, our method can be expanded to select the combination of patterns which differentiate in a fashion most similar to an outside criterion. The expanded method is called Differential Pattern Analysis.

"Suppose, for example, that our problem were to isolate the major patterns which would best differentiate fifty mental patients from fifty normals on a test of 100 items. In this case, we would proceed for each subject (patients and normals), in the same fashion as we did above for Subject 83; we would determine the major patterns for each patient in terms of other patients and for each normal in terms of other normals. This step would yield a set of patterns derived from patients and another set derived from normals.

"Using both patient patterns and normal patterns, we would compute the agreement score of every pattern with every other pattern and place them in a matrix. We would then select those patterns uniquely characteristic of either patients or normals. A pattern analysis of the matrix would facilitate this operation.

"A cross validation would be required to determine the ultimate value of the selected patterns for obtaining the desired differentiation between patients and normals ...

#### The Variability of Categories into which People Classify

"A further consideration of the above methods emphasizes an important and fundamental problem: The categories of persons with which any given person can be classified is a function of both the test items in terms of which the person is assessed and the group of persons with whom he is compared." (McQuitty, \*).<sup>2</sup>

#### Summary

To understand a single response by a single individual is a fundamental problem. This problem emphasizes that we must decide from theory or some other point of view both (a) with what other responses it is helpful to interpret the given response and (b) with what other individuals it is helpful to interpret the behavior of the given individual. The effectiveness of our interpretation of the single response

by the single individual depends on the validity of our choices in selecting other responses and other individuals with which to compare those of the single individual.

Once the above decisions have been consummated with high validity, then the methods of this paper and other similar methods are helpful in isolating meaningful personality structures.

Two especially effective methods of pattern-analysis are described in this paper. Both methods begin with a matrix of interassociations between people (or other objects). One method searches for internally consistent submatrices. These are usually small, each consisting of only a few individuals. They are initial indicators of statistical types, which are relatively hidden in the data. They are analyzed in a fashion which clarifies their appearance and develops them into larger types. Through this procedure a hierarchical classification of statistical types is constructed from the bottom up.

By way of contrast, Intercolumnar Correlational Analysis builds the hierarchical structure from top down. It divides a matrix into two or more submatrices, at least one of which represents a statistical type. It continues by dividing and redividing submatrices until at the bottom every person is represented as an individual type. This method has the advantage of using all indices of every matrix or submatrix in making its classifications, while crucial decisions in the above method are based primarily on only a few indices.

References

- McQuitty, L. L. Pattern Analysis--A Statistical Method for the Study of Types. In Chalmers, W.E., et al., Labor Management Relations in Illini City. Vol II Champaign, Illinois: Institute of Labor & Industrial Relations, University of Illinois, 1954, Chapter 14, 439-474.
- McQuitty, L. L. Differential Validity in Some Pattern Analytic Methods. In Bass, Bernard, M. and Berg, Irwin A. Objective Approaches to Personality Assessment. Van Nostrand Company, Inc., Princeton, New Jersey, 1959, Chapter IV, 66-82.
- McQuitty, L. L. Rank Order Typal Analysis. Educ. Psychol. Measmt., 1963, 23, No. 1 55-61.
- McQuitty, L. L. Capabilities and Improvements of Linkage Analysis as a Clustering Method. Educ. Psychol. Measmt., 1964, 24, No. 3, 441-456.
- McQuitty, L. L. A Conjunction of Rank Order Typal Analysis and Item Selection. Educ. Psychol. Measmt., 1965, 25, No. 4, 949-961.
- McQuitty, L. L. Single and Multiple Hierarchical Classification by Reciprocal Pairs and Rank Order Types. Educ. Psychol. Measmt., 1966 (a), 26, No. 2, 253-265.
- McQuitty, L. L. Multiple Rank Order Typal Analysis for the Isolation of Independent Types. Educ. Psychol. Measmt., 1966 (b) 26, No. 1, 3-11
- McQuitty, L. L. A Mutual Development of Some Typological Theories and Some Pattern-Analytic Methods, Educ. Psychol. Measmt., 1967, 27, No. 1
- \*McQuitty, L. L. Group Based Pattern Analysis of the Single Individual. Submitted for publication to Multivariate Behavioral Research.
- \*\*McQuitty, L. L., and Clark, J. A. Clusters from Iterative, Intercolumnar Correlational Analysis. Submitted for publication to Educ. Psychol. Measmt.
- \*\*\*McQuitty, L. L. A Novel Application of the Coefficient of Correlation in the Isolation of both Typal and Dimensional Constructs, Educ. Psychol. Measmt., in press.
- \*\*\*\*McQuitty, L. L. Multiple Clusters, Types, and Dimensions from Iterative Intercolumnar Correlational Analysis, submitted for publication to Multivariate Behavioral Research.
- McQuitty, L. L., Abeles, N., and Clark, J. A. A Search for Objective Syndromes of Psychopathology, study in progress.
- McQuitty, L. L., Clark, J. A., and Price, L. The Problem of Ties in a Pattern Analytic Method, Educ. Psychol. Measmt., 1967, 27, No. 4
- Schubert, G. A. Jackson's Judicial Philosophy: an Exploration in Value Analysis. Am. Pol. Sci. Rev., 1965, 59, 941-963.
- Zubin, J. A Technique for Measuring Like-Mindedness, J. of Ab. and Soc. Psychol. 1938, 33, 508-516.



Footnotes

- 1 Throughout this paper, quotations are included which refer to tables and figures. Whenever necessary, in order to make the code number correspond to the order in which tables and figures appear in this paper they have been changed within the quotation.
- 2 Appreciation is expressed to Multivariate Behavioral Research for permission to quote from McQuitty, Louis L., Group Based Pattern Analysis of the Single Individual, in press (Letter from the Editor, Dr. Desmond S. Cartwright, to the author, dated 16 February 1967)

McQuitty

Table 1  
Agreement Scores between Companies\*

	A	B	C	D	E	F	G	H
A		<u>29</u>	16	16	14	6	11	7
B	<u>29</u>		17	17	13	6	8	10
C	16	17		<u>26</u>	10	8	9	13
D	16	17	<u>26</u>		10	12	11	11
E	14	13	10	10		<u>21</u>	17	13
F	6	6	8	12	<u>21</u>		19	17
G	11	8	9	11	17	19		<u>24</u>
H	7	10	13	11	13	17	<u>24</u>	

\*Data from McQuitty, 1954

McQuitty

Table 2

Agreement Scores between Hierarchical Types

	AB	CD	EF	GH
AB		<u>13</u>	4	5
CD	<u>13</u>		4	6
EF	4	4		<u>10</u>
GH	5	6	<u>10</u>	

McQuitty

Table 3

Agreement Scores of Table 1 Converted to Ranks within Columns

	A	B	C	D	E	F	G	H
A	1	2	4	4	4	$7\frac{1}{2}$	$5\frac{1}{2}$	8
B	2	1	3	3	$5\frac{1}{2}$	$7\frac{1}{2}$	8	7
C	$3\frac{1}{2}$	$3\frac{1}{2}$	1	2	$7\frac{1}{2}$	6	7	$4\frac{1}{2}$
D	$3\frac{1}{2}$	$3\frac{1}{2}$	2	1	$7\frac{1}{2}$	5	$5\frac{1}{2}$	6
E	5	5	6	8	1	2	4	$4\frac{1}{2}$
F	8	8	8	5	2	1	3	3
G	6	7	7	$6\frac{1}{2}$	3	3	1	2
H	7	6	5	$6\frac{1}{2}$	$5\frac{1}{2}$	4	2	1

McQuitty

Table 4

Pearsonian Coefficients of Correlation between the Companies

Based on Raw Scores

A	B	C	D	E	F	G	H	
A	+1.0000	+0.8530	-0.0590	-0.1410	-0.3260	-0.4860	-0.4880	-0.5310
B	+0.8530	+1.0000	-0.1170	-0.0030	-0.4360	-0.5230	-0.5380	-0.4170
C	-0.0590	-0.1170	+1.0000	+0.5640	-0.2060	-0.4520	-0.3010	-0.2420
D	-0.1410	-0.0030	+0.5640	+1.0000	-0.3960	-0.2660	-0.2940	-0.2920
E	-0.3260	-0.4360	-0.2060	-0.3960	+1.0000	+0.4600	-0.0150	-0.0450
F	-0.4880	-0.5230	-0.4520	-0.2660	+0.4600	+1.0000	+0.1810	+0.0850
G	-0.4880	-0.5380	-0.3010	-0.2940	-0.0150	+0.1810	+1.0000	+0.4610
H	-0.5310	-0.4170	-0.2420	-0.2920	-0.0450	+0.0850	+0.4610	+1.0000

McQuitty

Table 5  
The Pearsonian Coefficients of Table 4 Converted  
to Ranks within Columns

	A	B	C	D	E	F	G	H
A	1	2	3	4	6	7	7	8
B	2	1	4	3	8	8	8	7
C	3	4	1	2	5	6	6	5
D	4	3	2	1	7	5	5	6
E	5	6	5	8	1	2	4	4
F	7	7	8	5	2	1	3	3
G	7	8	7	7	3	3	1	2
H	8	5	6	6	4	4	2	1

Table 6

Fifth Intercolumnar Matrix of Table 4

	A	B	C	D	E	F	G	H
A	+1	+1	+1	+1	-1	-1	-1	-1
B	+1	+1	+1	+1	-1	-1	-1	-1
C	+1	+1	+1	+1	-1	-1	-1	-1
D	+1	+1	+1	+1	-1	-1	-1	-1
E	-1	-1	-1	-1	+1	+1	+1	+1
F	-1	-1	-1	-1	+1	+1	+1	+1
G	-1	-1	-1	-1	+1	+1	+1	+1
H	-1	-1	-1	-1	+1	+1	+1	+1



Table 7

## The Emergence of Types AB and CD

A	B	C	D	A	B	C	D
A +1.0000 +0.8530 -0.0590 -0.1410				A +1.0000 +0.9696 -0.9122 -0.9587			
B +0.8530 +1.0000 -0.1170 -0.0030				B +0.9696 +1.0000 -0.9650 -0.8885			
C -0.0590 -0.1170 +1.0000 +0.5640				C -0.9122 -0.9650 +1.0000 +0.7632			
D -0.1410 -0.0030 +0.5640 +1.0000				D -0.9587 -0.8885 +0.7632 +1.0000			
Original Correlations				First Intercolumnar Matrix			
A	B	C	D	A	B	C	D
A +1.0000 +0.9987 -0.9936 -0.9970				A +1.0000 +1.0000 -1.0000 -1.0000			
B +0.9987 +1.0000 -0.9979 -0.9920				B +1.0000 +1.0000 -1.0000 -1.0000			
C -0.9936 -0.9979 +1.0000 +0.9819				C -1.0000 -1.0000 +1.0000 +0.9999			
D -0.9970 -0.9920 +0.9819 +1.0000				D -1.0000 -1.0000 +0.9999 +1.0000			
Second Intercolumnar Matrix				Third Intercolumnar Matrix			

Table 8  
Agreement Scores between Objects

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A		20	29	20	25	24	20	13	29	20	23	18	28	28	24	30	28	16	22	20
B	20		20	25	17	15	26	20	15	25	13	26	20	20	27	21	25	20	18	25
C	29	20		19	26	34	23	13	33	20	30	18	27	31	26	34	32	19	30	22
D	20	25	19		22	18	25	20	18	22	18	27	20	19	25	19	22	21	19	25
E	25	17	26	22		24	17	23	23	16	28	20	18	23	20	21	24	14	26	14
F	24	15	34	18	24		20	12	33	17	32	15	29	29	23	30	28	16	32	20
G	20	26	23	25	17	20		14	19	30	18	26	18	22	25	23	26	24	16	28
H	13	20	13	20	23	12	14		12	15	16	18	13	11	21	11	13	15	18	12
I	29	15	33	18	23	33	19	12		20	29	15	30	28	26	33	25	17	29	21
J	20	25	20	22	16	17	30	15	20		14	20	21	21	28	22	24	27	15	31
K	23	13	30	18	28	32	18	16	29	14		16	25	24	21	29	22	13	31	16
L	18	26	18	27	20	15	26	18	15	20	16		15	17	23	15	19	21	18	21
M	28	20	27	20	18	29	18	13	30	21	25	15		27	26	30	26	17	23	23
N	28	20	31	19	23	29	22	11	28	21	24	17	27		24	30	27	19	23	22
O	24	27	26	25	20	23	25	21	26	28	21	23	26	24		25	24	22	24	27
P	30	21	34	19	21	30	23	11	33	22	29	15	30	30	25		28	17	26	23
Q	28	25	32	22	24	28	26	13	25	24	22	19	26	27	24	28		18	22	22
R	16	20	19	21	14	16	24	15	17	27	13	21	17	19	22	17	18		14	30
S	22	18	30	19	26	32	16	18	29	15	31	18	23	23	24	26	22	14		16
T	20	25	22	25	14	20	28	12	21	31	16	21	23	22	27	23	22	30		

McQuitty

Table 9  
First Matrix of Intercolumnar Correlations of Table 8

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A +1.00	-0.25	+0.87	-0.42	+0.40	+0.85	-0.04	-0.29	+0.85	-0.08	+0.74	-0.47	+0.87	+0.93	+0.23	+0.90	+0.84	-0.19	+0.69	-0.02
B -0.25	+1.00	-0.45	+0.84	-0.48	-0.39	+0.76	+0.13	-0.31	+0.73	-0.42	+0.74	-0.31	-0.29	+0.51	-0.40	-0.10	+0.75	-0.65	+0.60
C +0.87	-0.45	+1.00	-0.51	+0.51	+0.95	-0.23	-0.35	+0.94	-0.19	+0.85	-0.57	+0.88	+0.92	+0.01	+0.93	+0.76	-0.40	+0.79	-0.16
D -0.42	+0.84	-0.51	+1.00	-0.62	-0.58	+0.63	+0.41	-0.56	+0.63	-0.61	+0.92	-0.51	-0.41	+0.22	-0.48	-0.26	+0.65	-0.61	+0.42
E +0.40	-0.48	+0.51	-0.62	+1.00	+0.60	-0.49	-0.00	+0.52	-0.66	+0.73	-0.47	+0.47	+0.35	-0.36	+0.46	+0.22	-0.61	+0.77	-0.50
F +0.85	-0.39	+0.95	-0.58	+0.60	+1.00	-0.32	-0.28	+0.95	-0.26	+0.92	-0.52	+0.81	+0.85	+0.05	+0.90	+0.64	-0.39	+0.87	-0.24
G -0.04	+0.76	-0.23	+0.63	-0.49	-0.32	+1.00	+0.01	-0.21	+0.86	-0.49	+0.55	+0.06	-0.01	+0.61	-0.12	+0.16	+0.84	-0.36	+0.85
H -0.29	+0.13	-0.35	+0.41	-0.00	-0.28	+0.01	+1.00	-0.33	-0.10	-0.25	+0.56	-0.45	-0.29	-0.42	-0.34	-0.25	-0.08	-0.19	+0.00
I +0.85	-0.31	+0.94	-0.56	+0.52	+0.95	-0.21	-0.33	+1.00	-0.25	+0.87	-0.54	+0.90	+0.92	+0.07	+0.94	+0.74	-0.33	+0.81	-0.14
J -0.08	+0.73	-0.19	+0.63	-0.66	-0.26	+0.86	-0.10	-0.25	+1.00	-0.39	+0.57	-0.04	-0.01	+0.63	-0.10	+0.09	+0.89	-0.45	+0.93
K +0.74	-0.42	+0.85	-0.61	+0.73	+0.52	-0.49	-0.25	+0.87	-0.39	+1.00	-0.58	+0.68	+0.74	-0.12	+0.73	+0.58	-0.46	+0.95	-0.38
L -0.47	+0.74	-0.57	+0.92	-0.47	-0.52	+0.55	+0.56	-0.54	+0.57	-0.58	+1.00	-0.41	-0.43	+0.17	-0.38	-0.18	+0.53	-0.60	+0.45
M +0.87	-0.31	+0.88	-0.51	+0.47	+0.81	+0.06	-0.45	+0.90	-0.04	+0.68	-0.41	+1.00	+0.91	+0.33	+0.93	+0.76	-0.13	+0.68	+0.03
N +0.93	-0.29	+0.92	-0.41	+0.35	+0.85	-0.01	-0.29	+0.92	-0.01	+0.74	-0.43	+0.91	+1.00	+0.27	+0.95	+0.90	-0.19	+0.69	+0.06
O +0.23	+0.51	+0.01	+0.22	-0.36	+0.05	+0.61	-0.42	+0.07	+0.63	-0.12	+0.17	+0.33	+0.27	+1.00	+0.30	+0.44	+0.61	-0.23	+0.67
P +0.90	-0.40	+0.93	-0.48	+0.46	+0.90	-0.12	-0.34	+0.94	-0.10	+0.73	-0.38	+0.93	+0.95	+0.30	+1.00	+0.83	-0.17	+0.70	-0.01
Q +0.84	-0.10	+0.76	-0.26	+0.22	+0.64	+0.16	-0.25	+0.74	+0.09	+0.58	-0.18	+0.76	+0.90	+0.44	+0.83	+1.00	+0.02	+0.52	+0.22
R -0.19	+0.75	-0.40	+0.65	-0.61	-0.39	+0.84	-0.08	-0.33	+0.89	-0.46	+0.53	-0.13	-0.19	+0.61	-0.17	+0.02	+1.00	-0.51	+0.89
S +0.69	-0.65	+0.79	-0.61	+0.77	+0.87	-0.36	-0.19	+0.81	-0.45	+0.95	-0.60	+0.68	+0.69	-0.23	+0.70	+0.52	-0.51	+1.00	-0.38
T -0.02	+0.60	-0.16	+0.42	-0.50	-0.24	+0.85	+0.00	-0.14	+0.93	-0.38	+0.45	+0.03	+0.06	+0.67	-0.01	+0.22	+0.89	-0.38	+1.00

Table 10

Fifth Matrix of Intercolumnar Correlations of Table 8

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A																				
B	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
C	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
D	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
E	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
F	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
G	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
H	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
I	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
J	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
K	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
L	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
M	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
N	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
O	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
P	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
Q	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
R	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
S	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0
T	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0	+1.0	-1.0

Table 11

A Portion of a Hypothetical N by N Matrix of  
Correlation Coefficients between People

	i	j	$x_1$	$x_2$	----- $x_N$
Both in the same type	i		$\bar{r}_{ix_1}$	$\bar{r}_{ix_2}$	
	j		$\bar{r}_{jx_1}$	$\bar{r}_{jx_2}$	
Each in any type	$x_1$	$*\bar{r}_{x_1i}$	$\bar{r}_{x_1y}$		
	$x_2$	$\bar{r}_{x_2i}$	$\bar{r}_{x_2y}$		
	----- $x_N$				

\*  $\bar{r}_{x_1i}$ , for example, denotes the mean of all  $r_{x_1i}$ 's.

McQuitty

Table 12

The Derivation of Pattern A-20 from those of Individuals A and 20

Items	1	2	3	4	5	6	7	8	9
Individual A	+	-	+	-	-	+	+	-	-
Individual 20	+	-	-	+	-	+	+	+	-
Pattern A-20	+	-			-	+	+		-

McQuitty

Table 13  
Hypothetical Data Illustrating the Computation of  
Agreement Scores between Patterns

	Code	2	3	7	8	16
Pattern 83-1		yes	yes	no	yes	yes
Pattern 83-2		yes	no	yes	yes	yes



McQuitty

Table 14  
Common Course Selection in Cluster 1, Figure 5

Intersection Points	
A	B
2	5
5	8
8	28
28	

Courses in Cluster 1, Figure 5:  
Personality, Experimental

2-General	8-Learning & Motivation
5-Personality	28-Abnormal

Table 15

Common Course Selection in Cluster 2, Figure 5

Intersection Points					
C	D	E	F	G	H
2	3	4	6	4	6
3	4	5		5	8
4	5	6		6	28
5	6			28	
6					

Courses in Cluster 2, Figure 5:  
Personality-Clinical, as Related to Business

2-General	6-Business & Personnel
3-Principles of Behavior	8-Learning & Motivation
4-Elem. Quan. Problems	28-Abnormal
5-Personality	29-Survey of Clinical

Table 16

Common Course Selection in Cluster 3, Figure 6

Intersection Points							
I	J	K	L	M	N	O	P
2	2	27	3	3	27	3	27
15	27	28	6	6		4	29
27	28		8	27		27	
28			27	28		29	
29			28				

Courses in Cluster 3, Figure 6:  
Individual Differences in Personality

2-General	8-Learning & Motivation
3-Principles of Behavior	15-Infancy & Preschool
4-Elem. Quan. Problems	27-Tests & Measurement
6-Business & Personnel	28-Abnormal
	29-Survey of Clinical

Table 17

Common Course Selection in Cluster 4, Figure 7

Intersection Points								
Q	R	S	T	U	V	W	X	Y
3	3	15	3	15	15	6	15	4
8	15	25	15		16	15	28	13
15	25		16		28	16		15
25			28			28		28
30								

Courses in Cluster 4, Figure 7:  
Dynamics of the Developing Individual

3-Principles of Behavior	15-Infancy & Preschool
4-Elem. Quan. Problems	16-Middle Childhood
6-Business & Personnel	25-Modern Viewpoints
8-Learning & Motivation	28-Abnormal
13-Social Movements	30-Dynamic Theories

Table 13

Common Course Selection in Cluster 5 Figures 8 &amp; 9

Intersection Points															
	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	NN	
3	3	3	3	2	3	2	2	2	3	3	3	3	3	3	3
4	8	8	8	3		3	3	3	5	5	28	4	4	5	
3	23	23		4		5	5	5	28	15		28	15	13	
16		28		5		15	23	8		28			28		
23				15		28		28					29		
				28											

Courses in Cluster 5, Figure 8 & 9: Personality-General

- |                          |                         |
|--------------------------|-------------------------|
| 2-General                | 13-Social Movement      |
| 3-Principles of Behavior | 15-Infancy & Preschool  |
| 4-Elem. Quan. Problems   | 16-Middle Childhood     |
| 5-Personality            | 23-Human Learning       |
| 8-Learning & Motivation  | 28-Abnormal             |
|                          | 29-Survey of Psychology |

Hierarchical Types

Imperfect Types

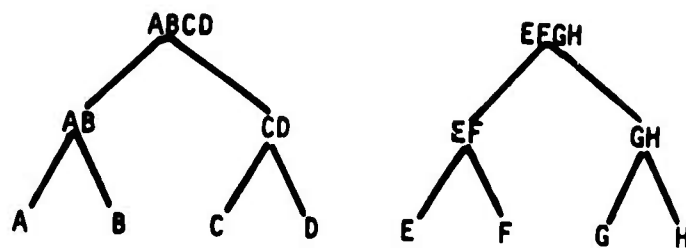


Fig. 1 Types of Companies in Terms of Some Union-Management Characteristics

McQuitty

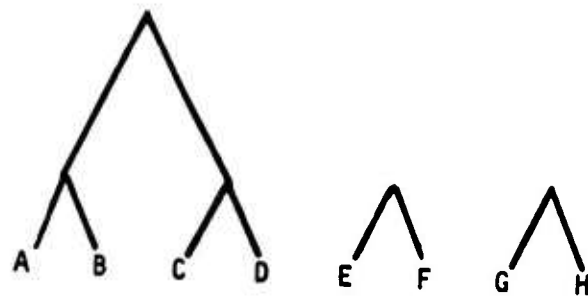


Fig. 2 Hierarchical Classification of Companies by Rank Order Typal Analysis of Agreement Scores

Agreement Scores, i.e., Number of Common Responses  
for the Silverware Patterns Grouped at each Intersection

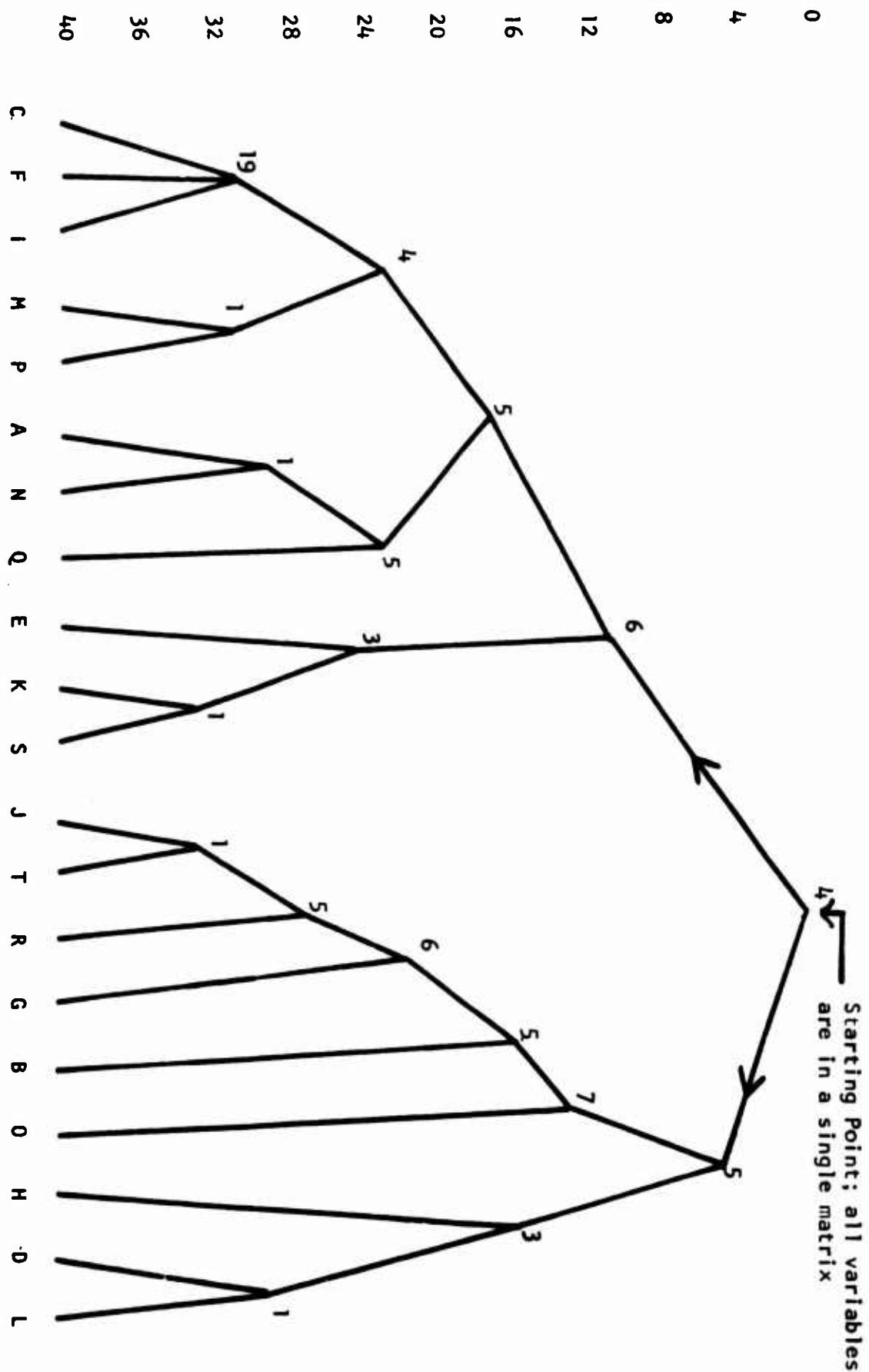


Fig. 3 A Typology of Certain Silverware Patterns  
Numbers at intersection points report number of iterations required to yield  
a matrix of plus and minus ones only



McQuitty

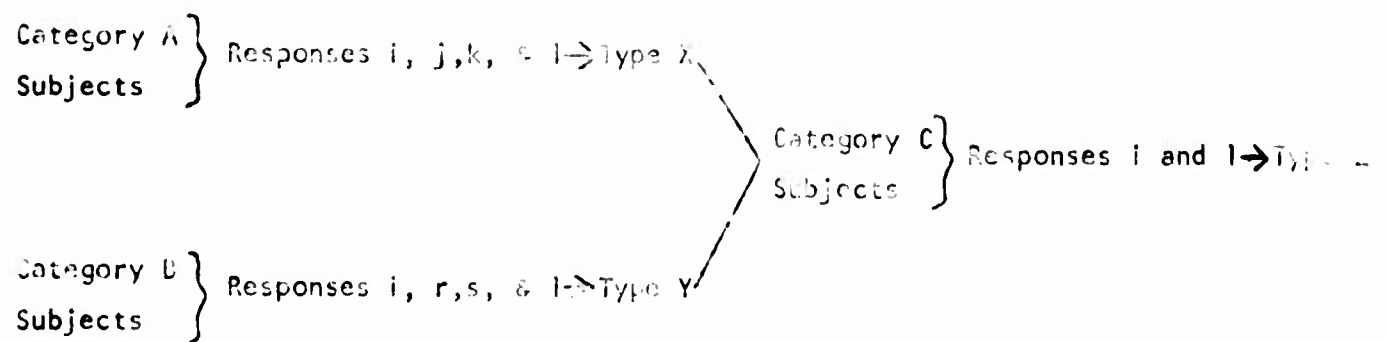


Fig. 4 Illustrating Two Kinds of Differential Validity

Number of Courses Taken by the Clusters of Students

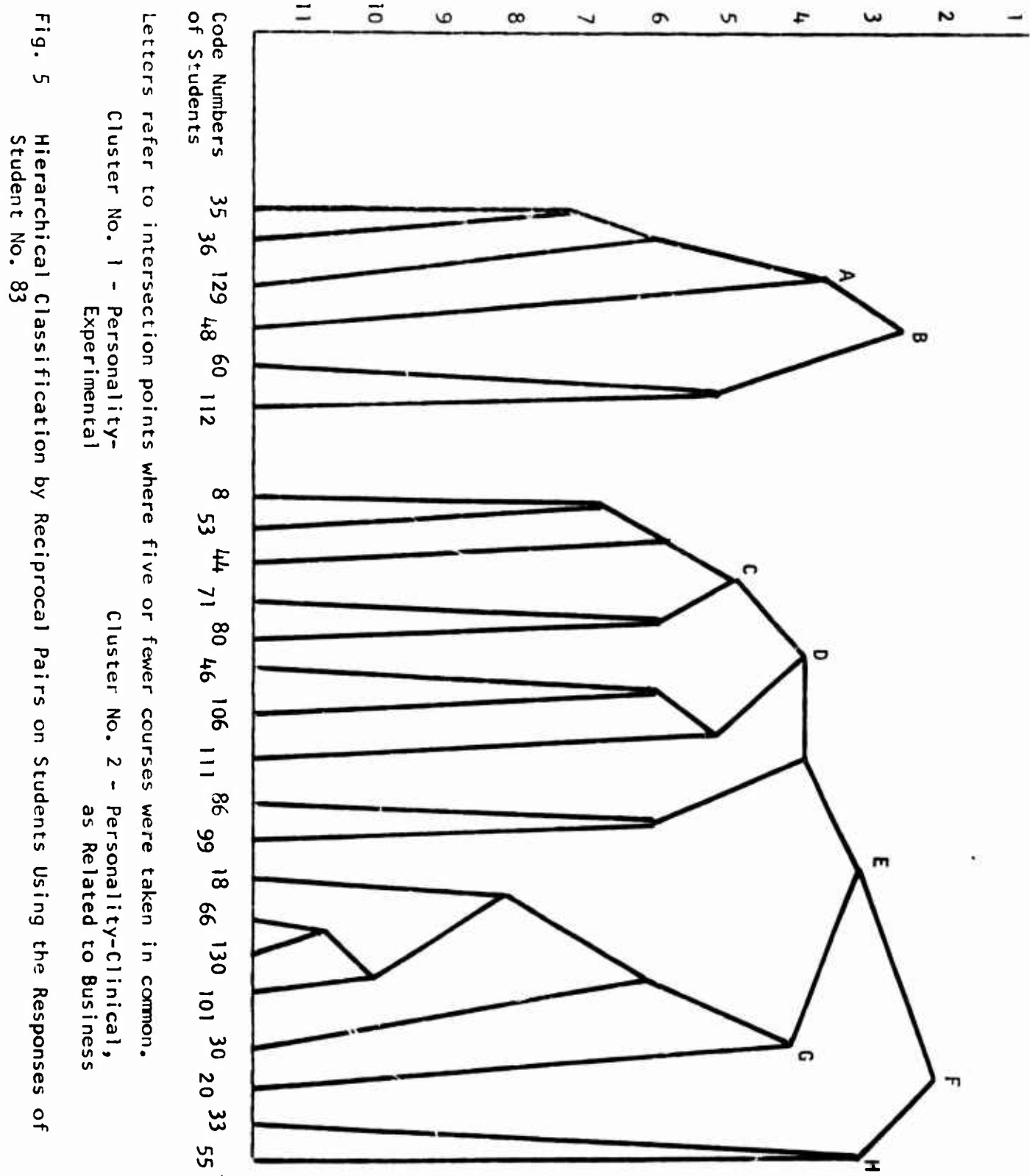


Fig. 5 Hierarchical Classification by Reciprocal Pairs on Students Using the Responses of Student No. 83

Number of Courses Taken by the Clusters of Students

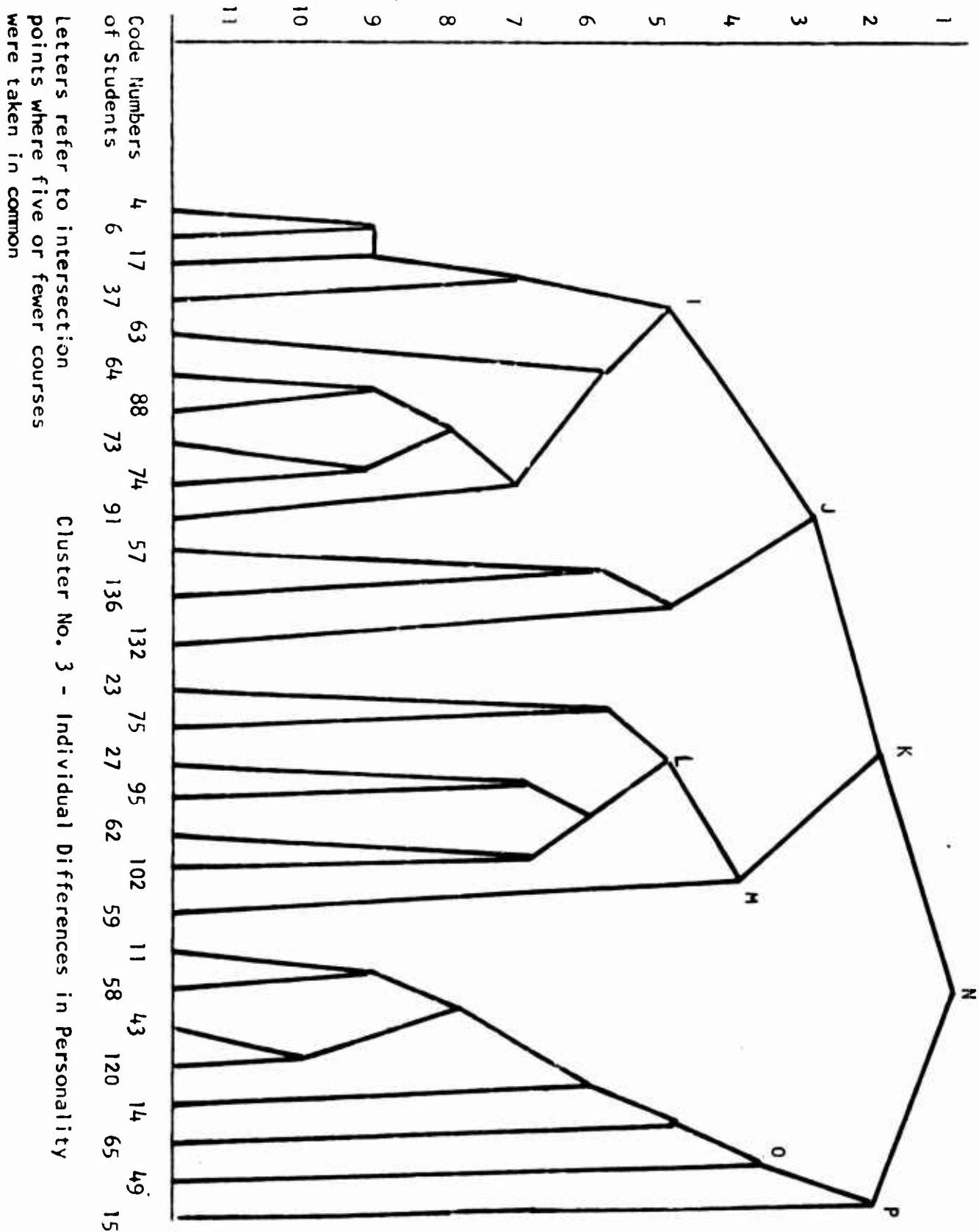


Fig. 6 Hierarchical Classification by Reciprocal Pairs of Students Using the Response of Student No. 83

Number of Courses Taken by the Clusters of Students

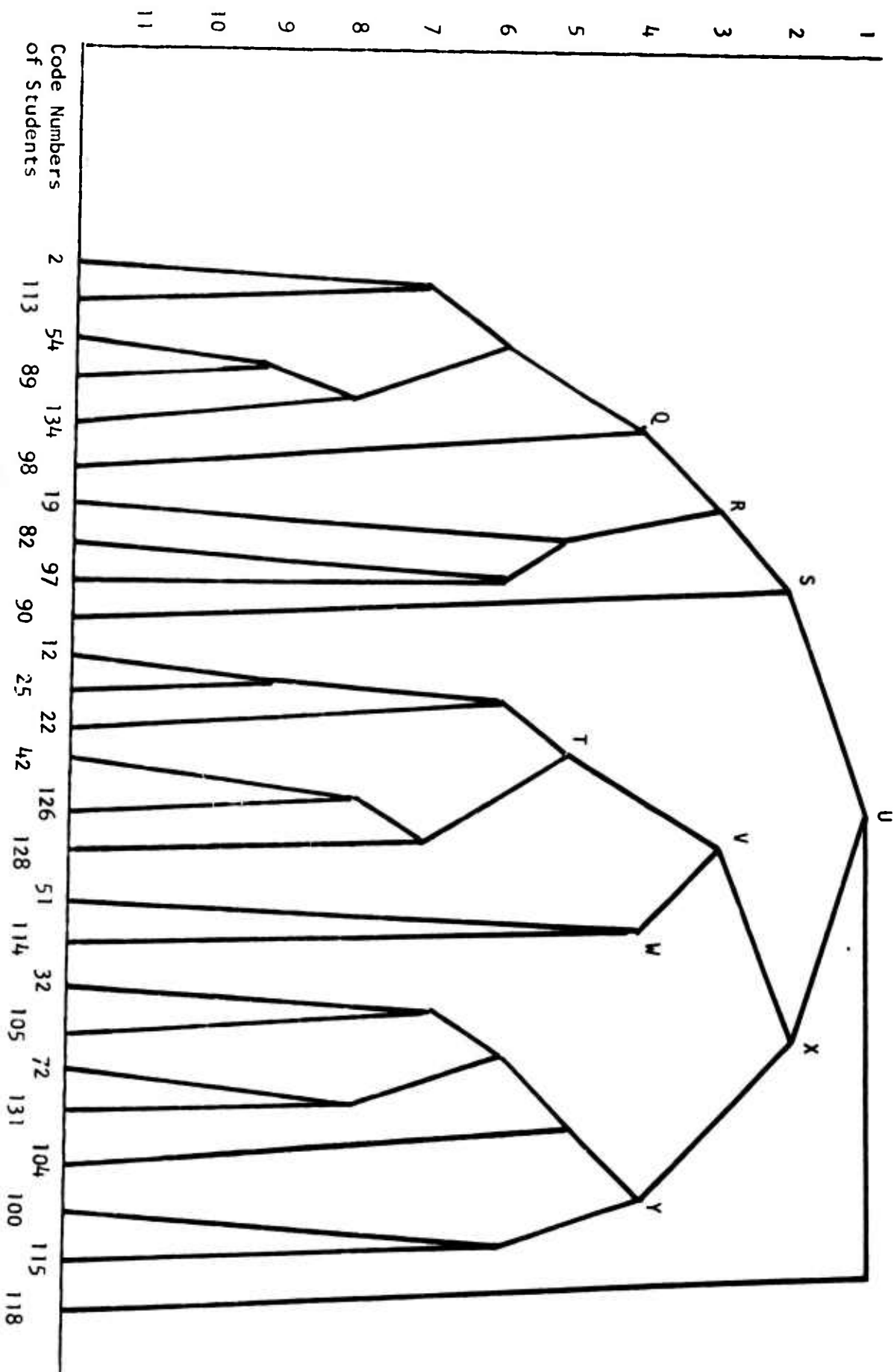


Fig. 7 Hierarchical Classification by Reciprocal Pairs on Students Using the Responses of Student No. 83

Letters refer to intersection points where five or fewer courses were taken in common

Cluster No. 4 - The Dynamics of the Developing Individual

Number of Courses Taken by the Clusters of Students

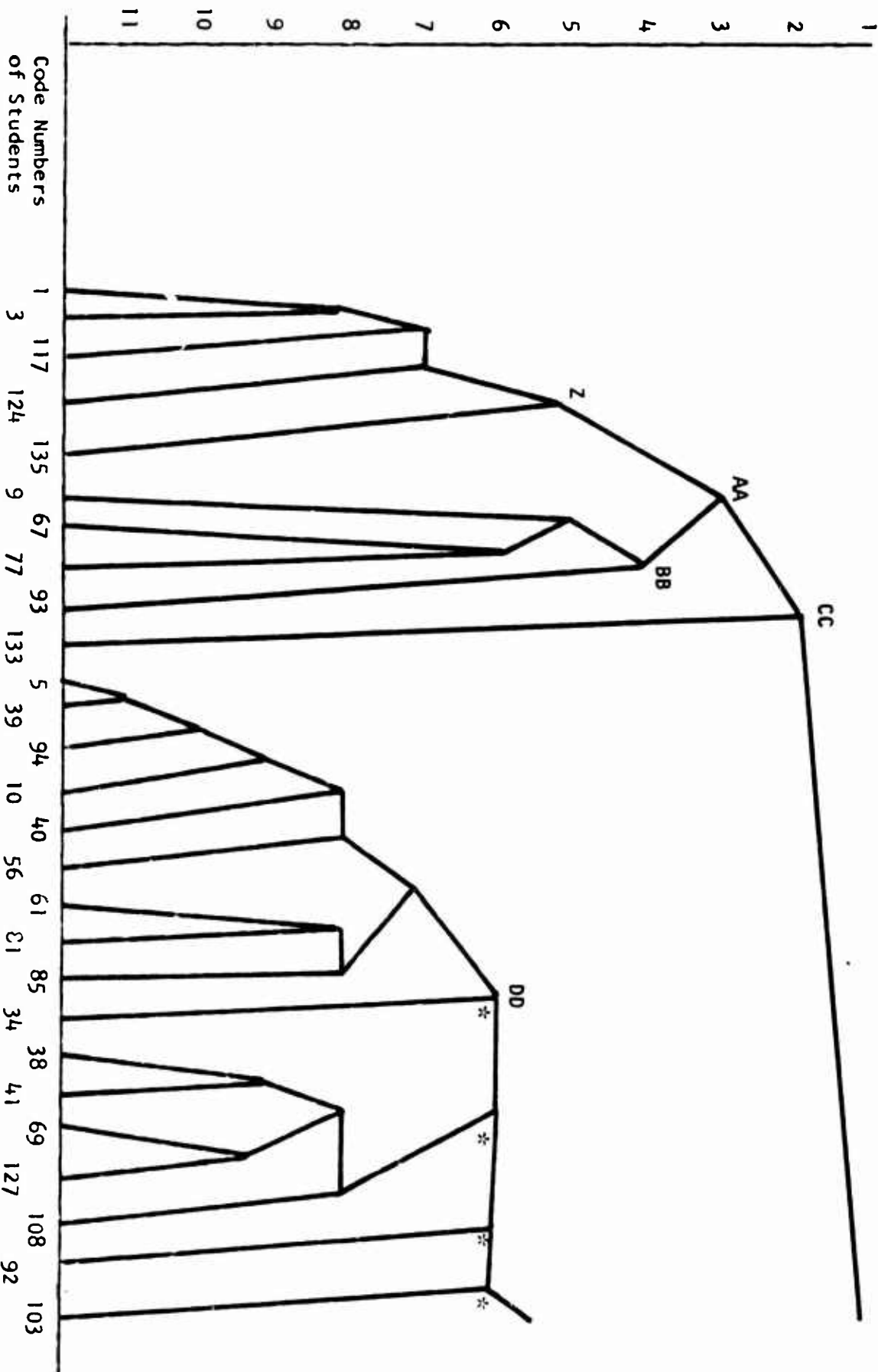


Fig. 8

Hierarchical Classification by Reciprocal Pairs on Students Using the Responses of Student No. 83 (Clusters of Courses Continued in Fig. 9)

Letters refer to intersection points where five or fewer courses were taken in common.

\*The above courses were taken by all subjects represented by all of these points

Number of Courses Taken by the Clusters of Students

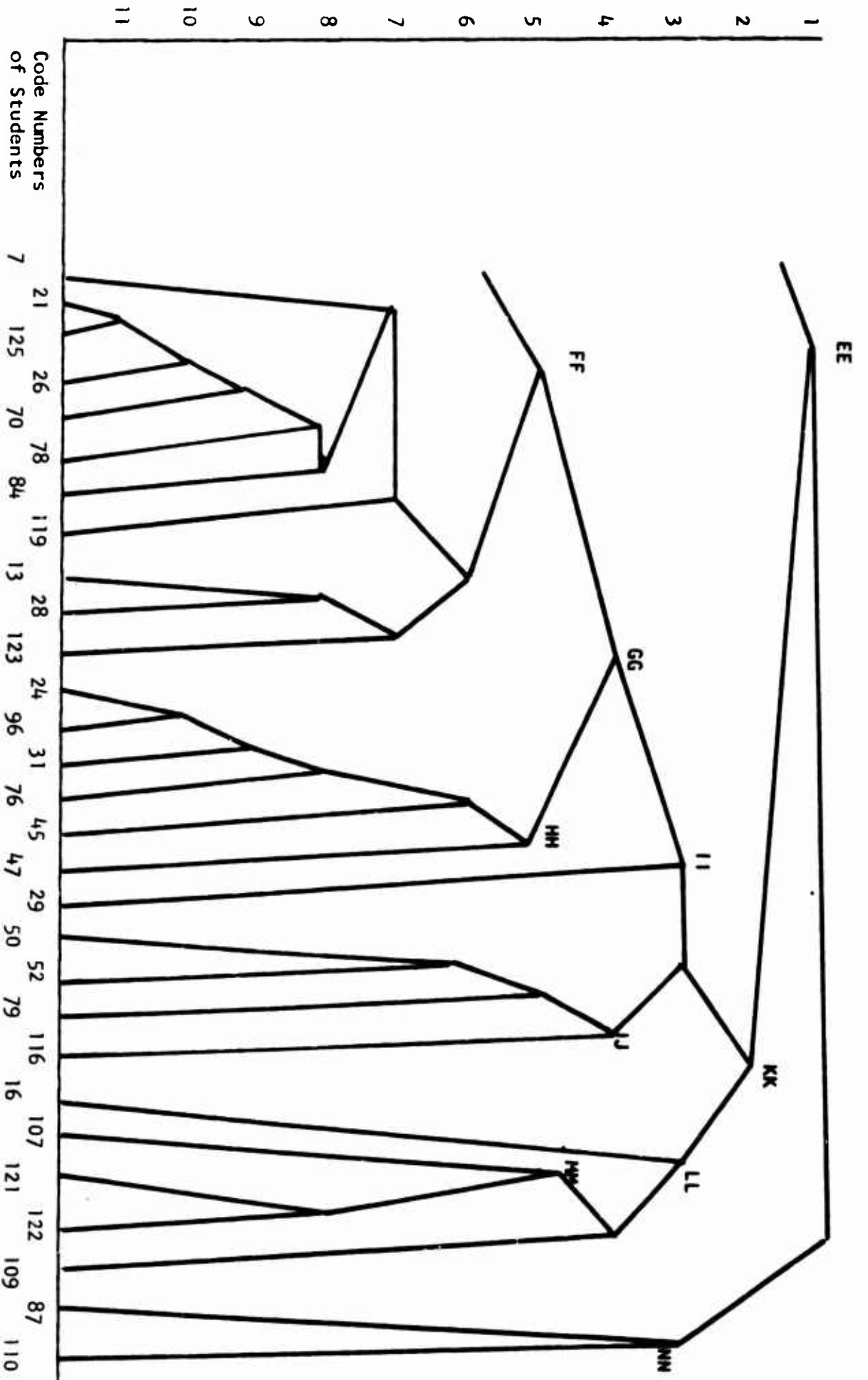


Fig. 9 Hierarchical Classification by Reciprocal Pairs on Students Using the Responses of Student No. 83 (Clusters of Courses, Continued)

Letters refer to intersection points where five or fewer courses were taken in common

Cluster No. 5 - Personality-General (Remainder)

# **Principles of Behavioral Taxonomy and the Mathematical**

## **Basis of the Taxonome Computer Program**

**Raymond B. Cattell**

**University of Illinois**

**and Malcolm A. Coulter**

**National Institute for Personnel Research  
Johannesburg, Union of South Africa**

### **1. Two Concepts of Type: Homostat and Segregate**

Placing people in types is an ancient pastime; but one still far from being fully understood in respect of both conceptual aims and methods of analysis. For example, the reciprocal relation of typing to the description and prediction by attributes and dimensions, discussed in the earlier Q-technique controversies (Burt, 1937; Cattell, 1951; Stephenson, 1936), yet remains to be properly worked out. To this day, the conceptual basis for types has remained crude compared to that developed clearly for attributes (by surface and source traits (Cattell, 1946), as defined in modern statistical models (Burt, 1950; Horst, 1965; Thurstone, 1947; Tucker, 1964)).

Elsewhere (Cattell, 1957), a list has been given of the rank and riotous verbal usages of "type". Such use as in Jung (1923), and many others who define types as the arbitrarily cut extremes of any bipolar continuous dimension, we shall set aside as more aptly handled by direct measurements on bipolar source traits. What we wish to designate as a type is the formal entity central to much psychological and biological classification, embodied in the last by the concept of a taxon, e.g., species, genus, family, etc. (In psychology, we need not necessarily adopt the biologist's further concern with "dendrograms," i.e., the arrangement in classificatory hierarchies.) Types appear in psychology as groupings by occupational skill, complexes of attitude in political groups, pathological syndromes, and by certain genetically determined patterns of behaviour.

Psychometrics has, in its main developments, ignored this granulation of its populations in favor of a simplified world of homogeneous normal distributions of characteristics and linear relations between them. Over the normal ranges of behaviour, the approximation has been good enough to permit the effective prediction of individual difference by means of broad personality factors. But as research broadens, the realities of more complex natural distributions demand to be considered. Considerations of efficiency require that our models begin more explicitly to encompass types, and the non-linear relations and pattern effects which go with them.

We shall, therefore, begin with the central, if initially over-simplified, definition of a type as the most representative pattern in a group of individuals

located by a high relative frequency -- a mode -- in the distribution of persons in multidimensional space. This definition will be made more stringent as we proceed. The principle possibilities are illustrated for one and two dimensions by Figure 1.

(Insert Figure 1 here)

Figure 1 is intended to bring out that: (1) non-normal, multimodal groupings can easily exist in a multivariate distribution even when the distribution projected on any one of the dimensions is virtually normal; (2) all modes are relative as to density, so that, as at  $A_1, A_2; B_1, B_2; C_1, C_2;$  and  $D_1, D_2$  in Figure 1, one can have "types within types"; and (3) that there are really two distinct possible definitions of type, one hinging on (a) high mutual similarity of members, i.e., all coming within a circumscribed distance of one another as illustrated by those lying in the dotted circles 1, 2, 3, and 4, and (b) forming part of a group in which, though some members may extend to remote distances from others, each is less remote from another member of that group than from individuals outside the group, e.g., as shown by the types B and C. Thus, persons in the regions A,  $C_1$ , and  $B_3$  constitute two types, 1 and 2, according to definition (a), whereas they fall into three types, A,  $C_1$ , and  $B_3$ , according to definition (b).

A definition with Euclidean or Boolean rigour for these two concepts will be given later, but on the temporarily adequate basis already given, we shall refer to them respectively by the term homostat, meaning "a set of people standing at closely similar positions in space", and segregate, implying "a set consisting of people continuously related through other people in the set and isolated from those outside, but not necessarily similar in position, i.e., not of high homogeneity". Readers may find it convenient, as we have in our own laboratory discussions, to designate them "stat" and "ait" respectively.

A glance at past psychological work on types e.g., McQuitty's pattern analysis (1963), that of Munnally (1962), Overall (1964), and of ourselves with the pattern similarity coefficient (1949, 1950, 1952), shows that attention has hitherto been operationally directed exclusively to homostats, despite the concept of segregates having sometimes been obviously present in the writer's mind.

## 2. The Most Promising Model, From a Scientific Standpoint

The main aims in research on types are: (1) To produce a methodology for operationally locating and identifying segregates and homostats. (2) To develop mathematico-statistical formulae, based on improved models of type, for utilizing test results for predictive purposes and for investigating laws which may arise from the peculiar nature of types.

Briefly to anticipate what this second step may comprise, we would point out that Aristotelian classification permits predictions of the kind: "This is a dog; therefore it may bite"; or "This is a schizophrenic; therefore the prospect of remissions is not high." In other words, a classification by variables of one kind may permit prediction on others not included in the immediate observations. As will be brought out later, the use of types need not



stop with this Aristotelean, categorical formulation. It will lead rather to the recognition that in numerical data, the relation of a "test" to a "criterion" may be very different within species from that obtaining between species. Thus, the relation of two variables could be non-linear across all individuals in the total genus, yet exactly linear within each species. The use of distinct in- and between-type dimensions instead simply of a single set of broad dimensions across a genus demands that before data is fed to the computer, one has to consult an encyclopedia (to recognize, by appropriate properties, each individual's belonging to a particular species). The reward, however, of this classificatory labour is likely to be a more accurate prediction from the individual's scores, or the discovery of clearer laws for the segregated types, obscured in the mixture of species in the genus.

As we proceed to more precise concepts for both discovering and using types, it is necessary (since particular exemplifications in, for example, zoology, psychology, astronomy, mineralogy are likely to differ) to define the breadth of our approach. Our aim is to be comprehensive (our association with Sokal and Sneath (1963) and Eades (1964) in applications to entomology has been encouraging and enlightening in this respect) and we believe that the psychologist, before he devotes his ingenuity to statistics, would do well to take a philosophical pause, for he needs to develop a plausible scientific (not merely a statistical) model of types, based on speculation as to how and why they arise. Briefly, our theory is that types arise from three causes: (1) Adaptive Success, because of special value in the combination (survival value in biology, utility of human artefacts), (2) Combinations Required by Natural Law, where a pattern repeats itself modally because it is required by a particular combination of natural laws, e.g., crystalline forms, cloud types, solar systems, and (3) Bio-social Gravitation. This supposes that once the beginnings of a type exist there will be a tendency by imitation for individuals to gravitate towards its centroid. This occurs socially in fashions and fads and biologically in species formation. (Sewall Wright's "genetic drift" has relations to the latter.) Obviously, psychology has types of all three kinds: the skill and personality patterns of different occupations are examples of functional adaptations; the behaviour pattern of delirium tremens or Huntington's chorea have no adaptive value and occur simply as inevitable patterns from laws of neurological breakdown, etc., cultural and racial types relate to the third source.

All three sources indicated by this theory of type origins would result in some combinations of parameters being represented by high (modal) frequencies while other zones (combinations) in the coordinate system, which theoretically might be filled, remain empty of individuals. Parenthetically, it will be mentioned that in functional adaptations dependent on either evolution or human invention, the additional possibility must be considered that some zones are unoccupied not because they necessarily represent a non-functional combination, but because for some reason they cannot be, or have not yet been reached. In biology, the intermediate mutational steps necessary to reaching some advantageous end pattern may be chemically unstable or biologically lethal. The giraffe's neck had time to grow gradually as his forelegs grew, so that he achieved the advantages of height without losing his capacity to drink; but other useful biological combinations might be too much of a "tour de force."

The matrix of scientific necessities out of which types are born will presumably be indicated in some degree by the varying textures, dendrograms (hierarchies) and cluster sizes emerging, as discussed in Section 8 below. However, both for adaptive types and natural law types, there is reason to expect that: (1) there will arise an unusually high frequency of cases in which some particular range of scores on parameter  $x$  is associated with a special range on parameter  $y$ , because this is functionally useful and is preserved and multiplied. Secondary pairs of optimum ranges will generally also exist, but apart from these modes, instances of individuals with other combinations will be rare; (2) among individuals at these modes some entirely new organs and therefore dimensions may appear which are not present in the general "population" (and, therefore, the distribution of these for the general population would have an extensive positive skew. For example, among the types on the tea table, only teapots distribute on the "length of spout" variable); (3) a class which we may call "across species" variables may be practically normally distributed over members of the whole genus despite many "species type" segregations, while the class of "within species" variables will, as stated, be badly skewed. These type concepts thus imply the recognition of three classes of variables, with greatest relevance, respectively within species, between species, and across the whole population of the genus; and (4) by reason of information in these variables, one will in general expect to have to complete the description of type segregation and distribution by reference to "higher order" structures, briefly indicated here by the terms textures and hierarchies.

Some slowness in coming to grips with the necessary concepts in this field must probably be ascribed to certain habits of mind, which favour simplified mathematical abstractions even when they fail to describe and do honour to the intrinsic irregularity of the data. Analytical geometers are not easily at home with topologists, and here even topologists themselves are being forced to face the intractable specificity of detail elsewhere faced only by topographers! The problem is very close to that of describing the actual cloud masses at a given moment in an  $n$ -dimensional sky. Even when this goal is admitted, most people begin by thinking of discrete cululus clouds neatly spaced in a summer sky, but are forced at the end to come to terms with the ultimate in irregular masses -- an October storm-wrack. Those who develop geometrical models and statistical procedures have no alternative but to brace themselves for this degree of complexity if they wish to describe the variety of human beings in a society or what actually happens in biological evolution.

### 3. Two Alternative Principles for Locating Stats

Anyone who has followed the history of psychologists' attempts to handle the type concept, with  $Q$ -sort,  $D^2$ , discriminant functions, Holzinger's  $B$  coefficient latent class analysis, etc., must admit that little of theoretical or practical psychological importance has yet emerged, and he may justifiably wonder whether the tools and concepts have been adequate. For example, psychiatrist's syndrome groupings, despite some application of correlation methods by Degan (1952), Huffman (see Cattell), Lorr (1962), Wittenborn (1951), and others continue to take their authority from subjective clinical impressions, while in social psychology and related areas, it is hard to point to any theory which has arisen from a statistically adequate demonstration of types.

Elsewhere (Cattell: 1951, 1952a), appeals have been made for recognition that in classifying individuals by resemblance methods: (1) Q-sort<sup>1</sup> is vitiated

---

<sup>1</sup>Q-sort, principally propagated by Rogers (1951), is a rank correlation version of what may be called a Q-bar ( $\bar{Q}$ -) technique. It is important clearly to distinguish Q-technique (sometimes called Q<sup>1</sup>-technique) which stops at finding correlation clusters, from true Q-technique, which is a full factor analytic technique (the transpose of R-technique) aimed at obtaining dimensions. Since Q-technique depends on the correlation coefficient, one cannot, for the above reasons, agree with its otherwise careful and precise use in the extensive taxonomic work of Sokal and Sneath (1963). Types are not factors.  $\bar{Q}$ -technique, on the other hand, yields types and will do so without throwing away important evidence if it uses  $r$  instead of  $r$ . Incidentally, for brevity we shall refer to the square matrix,  $P$ , with the same people at top and side, which is a common beginning of all the above resemblance methods of typing, as a "Q matrix."

---

so long as variables rather than factors are used, and without a principle for sampling variables. For indices of resemblance are completely unstable and meaningless without either resolving variables into factors or taking them in a stratified sample; (2) use of the correlation coefficient gives misleading results, for it throws away indispensable information, recording only the shape similarity of two profiles without reference to level or accentuation (Cattell, 1951); (3) Holsinger's B coefficient (Holsinger and Harman, 1941) disregards the difference between nuclear and phenomenal cluster structure which is discussed below; (4) latent class (sometimes called latent "structure") (Lasarsfeld, 1960), though a statistically clearly developed method, does not meet the need for a parametric treatment of the assignment of individuals to classes; (5) the multiple discriminant function is not a means of finding types, but only of giving emphasis, rigidity, and apparent precision to groupings initially discovered by other and usually more subjective methods.

Faced with these inadequacies of present type concepts and search methods, one recognizes the possible need for a radical re-orientation. This is reached first by realizing that the idea of type hides two very distinct concepts -- entities which we call stats and aits -- and, secondly, by recognizing two distinct methodological approaches to locating these in nature. The implicit definition of a stat (or "homostat") above can be sharpened now to a set of individuals within which the mutual resemblance of all pairs exceeds a certain value, significantly higher than that obtaining between pairs in the population at random. Although a segregate (or "ait") is different, we shall find that the stat is a necessary concept in reaching it, so the location of stats is first treated here and the operational definition of aits is deferred.

To locate a stat, one of two broadly different approaches are open to us, as follows:

(1) The Inter-Id Relation Method. This starts with people (or other individual patterns)<sup>2</sup> as reference points in a space defined by coordinates corresponding to the factors, etc., by which individuals are measured.

<sup>2</sup>Some general term is badly needed for the individuals who provide us with the dimensions (in factor analysis) or types ultimately extracted. Although the psychologist commonly thinks of people, these entities (each defined as a pattern), even in psychology, must also include such things as groups, processes, culture patterns, etc. Since a set of firmly defined and inter-related terms for all elements in the basic data relation matrix has been adopted in the new Handbook of Multivariate Experimental Psychology (Cattell, 1966), I am proposing to use here consistently the term id for any and all entities of this kind. That is to say, a Q-matrix is defined as bordered by ids and having in its cells scalar quantities expressing the relation between ids. Incidentally, this usage of id is so remote from the other usage in psychology, if psychoanalysis is to be so, that no comparison can exist.

It calculates the distance of each person from every other, locating first the dense "plexuses" of people, and secondarily, the position of the centroids of such groups. Most simply, a square matrix (a "Q matrix") is set up bordered by the same set of people on the two sides. Into the cells are entered the quantities which express the similarities of the members of the pairs defined each by a row and a column. Methods can then be developed to find the clusters of people constituting stats, and later, aits.

(2) The Density-in-Space or "Cartet Count" Method. Here, one begins with ids placed in position in a coordinate system by a matrix of scores on the coordinates. Convenient intervals are then taken on these coordinates to define "cartets" -- which, in a two-dimensional map, would be squares fixed by boundaries of latitude and of longitude. A computer program can then be written to count the number of ids in each such square ("hyper-cube" or, most generally, "cartet," if we may suggest such a term (after Descartes)) for such a rectangular subspace in a lattice of Cartesian products. Fixing a "significantly high density" count by relation to the average total density, one could first set aside stats, and, by secondary process discussed below, aits. Some experiments would be necessary regarding the size of the component subsets of Cartesian products in order to best bring out the modal groupings in relation to the general texture of the domain.

One must recognize from the beginning, however, that the "cartet-count" method will soon reach a number of cartets to be counted that could be onerous even for an electronic computer. The difficulty is illustrated by the fact that with only 16 dimensions, and intervals restricted to just 6 in number (3 plus and 3 minus) subtending each one standard score, for each coordinate the total number of hypercubes (cartets) which would have to have their contents examined by counting is  $6^{16} = 2,830,000,000,000$  (approx.). On the other hand, the number of resemblance entries to be examined in a Q-matrix, by the inter-id method ((1) above), also increases exponentially and is fairly formidable with four or five hundred people. Typing procedures, with any adequate sample of ids, are necessarily and characteristically going to be demanding of computer time. In practice, with the cartet count approach, one might often be content to use only two coarse score intervals per coordinate scale, but in a 16 element profile (which is probably fairly typical of psychological needs),

the cube count would still have to cover  $2^{16} = 240,000$  counts.

In this article, we shall concentrate on the inter-id approach since the cartet count procedure is obvious. Here, one needs first to find a suitable index of resemblance between any two ids, presuming each to be already measured on a profile of dimensions. For this, and allied purposes, the profile similarity index,  $r_p$ , has been developed (Cattell, 1949) as being free of the drawbacks of  $r$ , of Mahalanobis'  $D$  (1936), and of some other at times popular indices.

The profile similarity coefficient,  $r_p$ , has the formula:

$$r_p = \frac{2k_m - \sum d^2}{2k_m + \sum d^2} \quad (1)$$

where  $k$  is the number of profile elements; each  $d$  is the difference, in standard scores, of the two people concerned, on any one profile element, and  $k_m$  is the median  $X^2$  value for  $k$  degrees of freedom. At  $k=20$ ,  $k_m=19.337$ , so that above, say, 20 profile elements there is not much argument for using  $k_m$  instead of a simple  $k$  in the first part of the numerator and denominator. The former will exactly divide the possible  $r_p$ 's into equal numbers of positive and negative values, but the former will give a zero sum of negative and positive  $r_p$ 's. The advantages of this  $r_p$  over the Mahalanobis (1936) distance function,  $D$ , are:

(1) That it gives comparable values from study to study in comparing two ids, regardless of the different metrics and numbers of profile elements. This it does because (a) all coordinate values are in standard score, not different units for each, and (b) the formula allows for differences in the number of coordinates (profile elements) in evaluating the "distance." Moreover, it behaves very similarly to the familiar correlation coefficient, registering 0 where the relation between two people is no better than chance, +1.0 when they are perfectly alike, and -1.0 where they are as unlike as possible. By contrast, one never knows what the meaning of a particular  $D$  value is without an elaborate consideration of additional circumstantial facts or the making of additional calculations.

(2) Since different investigations in the same domain often differ somewhat in the number of dimensions they employ, both of the above features (1(a) and 1(b)) help in surveys attempting to integrate conclusions about types.

(3) A significance test has been worked out by Horn (1961) for  $r_p$  and other properties are under investigation. This and other developments of the index promise reasonable prospects of programs making it negotiable in further areas.

On the other hand, unlike  $D$ ,  $r_p$  does not have simple Euclidean distance properties. The relation of  $r_p$  to  $D$  (when the latter is put in standard scores form) is shown in Figure 2. It will be seen that in the central range, it is

approximately linear and the relation is still closer to linearity for the special  $r_p$  derivative,  $r_n$ , proposed below. However, non-Euclidean spaces, if properly understood, can be as usefully manipulated as Euclidean and the temptation of convenience offered by the use of familiar space must be rejected if the Euclidean representation,  $D$ , does not also give the greater psychological convenience, e.g., in vocational selection, etc., which is provided by the use of  $r_p$ .

(Insert Figure 2 here)

With this crystallization of an acceptable means by which the similarity of ids, i.e., of people, stimulus situations, groups, processes,<sup>3</sup> etc.,

---

<sup>3</sup>Parenthetically, to ward off any incorrect assumption of formal narrowness in our approach, let it be noted that the whole treatment of similarity by attributes as proposed here includes application to processes as well as structures. A psychiatrist, for example, may say that his assignment of an individual to the syndrome type "schizophrenia" includes observations on the course of onset itself, and the notion of a malign outcome. Such process attributes can, of course, be included along with structured, "immediate" measures in the designation of a specific profile of measures. (When  $r_p$  is used thus to locate types of processes rather than types of persons, certain time sequence information, distinguishing a configuration (Cattell, 1957, p. 396) from a profile, must be included.) The procedure can also be used for grouping processes as such, as discussed in detail elsewhere (Cattell, 1966).

---

measured can be measured, using a profile of dimensions, let us turn to the next problem. This concerns the use of such an index in the id-relation procedure for finding types.

#### 4. Defining Stats in (a) General Purpose Dimensions

##### and (b) Special Criterion Functions

It is part of the conceptual inadequacy of the approaches hitherto made by scientists to the type concept -- even in some of the best technical work, as of that by McQuitty (1963) or Sokal and Sneath (1963) -- that operations have been set up to find a homostat without recognizing that it will not have any uniqueness of center and boundary, for such uniqueness is characteristic only of a segregate. This arbitrariness and subjectivity of the stat, not only in width but also in position and even with an exact index of similarity, can be most quickly realized by a two dimensional example, as in Figure 1.

The investigator has to begin with some choice of similarity level as "significant" or "outstanding"; and though a rationale is given below for computing a finally objective boundary value for  $r_p$ , the limit of belonging must initially be arbitrary and tentative. Defining a stat by the property that every individual in it must resemble every other above this limit means that in Figure 1 (neglecting the slight departure of  $r_p$  from an Euclidean



distance) all people within a circle of diameter equal to  $r_p$  are in the same stat. If the  $r_p$  limit is well chosen, the majority of such circles drawn over much of the graph will each enclose less than two people. Only where there is a dense modal accumulation will more than two people, i.e., a type, be "lassoed" as, for example, at 1, 2, 3, and 4.

However, we must note (a) that sometimes, as at 2, individuals in two distinct aits will be caught in one stat, and (b) that, as in 3 and 4, two correctly defined stats will nevertheless overlap. In fact, there could be a whole series of stats in a dense area, each including many ids and different only in the inclusion of one different person from its neighboring stat. This can readily be seen if we imagine, say, 50 persons evenly following in the given coordinate space in a long row, with a circle diameter chosen to include, say, 3 persons at a time and finishing with 48 3-stats.

Furthermore, when we pursue in Section 6 below, the operational steps for locating stats from Q matrices filled with similarity indices, i.e., for going from the algebraic to the geometrical view here briefly invoked, it will be found that the psychologist needs two distinct concepts of stats, which are there named and defined as nuclear and phenomenal stats. Nevertheless, despite these complications and uncertainties of ultimate inference, the recording of stats has both direct value in itself and ancillary value also in providing a basis for proceeding to aits. With this foretaste of the problem of discovering stats, to which we shall return, we must pause a moment to solve a prior problem, one which stands squarely in the way of our progress, namely, an uncertainty about the very meaning of similarity. For in spite of the apparent precision of our  $r_p$  coefficient of profile similarity, it will become apparent that we cannot use it in all situations we might encounter until we have corrected it to less restrictive assumptions. In fact, we must pause briefly in this section to make some almost philosophical inquiries about the purpose and setting of its use.

The design of  $r_p$  has cleared it of giving accidental and unknown weights to different profile constituents, but it has left it with the rigid assumption that all dimensions receive exactly equal nominal weight -- and this may not fit all purposes.

As was pointed out in the original logic for  $r_p$  (Cattell, 1949) both the philosopher and the man in the street have always been haunted by a distinction between the character of the object in itself ("das Ding an sich" of Kant) and what it does or is useful for. (Perhaps even Hume's "primary" and "secondary," or the theologian's "grace" and "works" might be related to this distinction.) Certainly in the operation of psychological prediction, we constantly and confidently make a distinction between traits or "predictors" and predicted performances or "criteria." Viewing this from the standpoint of type distributions, it is easy to see that the modal groupings we would get on certain criteria will differ from those we would get on the total profile of traits. By the same token, for a barber, a brush, a comb, and a pair of scissors belong in the same class, while a screwdriver and a bottle opener do not. But by the urgent drinker, the bottle opener, the screwdriver and the scissors are seen as promising members of a class to which brush and comb do not belong.

The present writers, in basic personality research, do not accept the psychometrist's differentiation of test and criterion as having any fundamental status. But the difference between the total personality profile of behaviour extracted traits and a single "criterion" (or any other) performance is real enough. The latter is always some quite specific mathematical function of the former. On such a specific function, the modal stats (or sits) will be peculiar to that derived variable, i.e., different from the distribution on other functions or in the  $k$ -dimensional trait profile space. This is the mathematical expression of the statement that "all classifications are subjective, depending on the purpose of the classifier." Indeed, it is the general profile classification which now begins to look doubtful and subjective compared to that on the concrete criterion and we find ourselves asking, "What do we mean when we talk about 'the thing in itself'?" By what right, for example, did we start by giving equal weight to measures on the  $k$  dimensions of the profile in the  $r_p$  calculation. The fact that our initial concept of shape comes from the physical world (Newtonian, at that) fools us in the wider context, for we are naturally accustomed to giving equal weight to height, length, and breadth measures. What the psychologist really has to deal with is a severe case of Einstein's world, with dimensions variously and severely contractible.

The only firm basis for a system of weights for dimensions, as pointed out by Burt (1937), by the present writer (1946, 1957) and by Kaiser and Caffrey (1965), is a concept of a population or universe of behavioural variables, from which the dimensions derive. A rigorous and operational basis for dimension weights in the personality realm has long been available in Cattell's personality sphere concept (Cattell, 1946; Cattell and Warburton, 1967). Employing equal weights for the  $k$  elements of the profile used in  $r_p$  is therefore justified only if one has demonstrated that these dimensions approach a certain relation to the personality sphere. That relation is that the squared loadings of each factor over the personality sphere of variables sum to the same value as for all other factors (or, since in practice one must work with a random or a stratified, that they approach equality within the limits of error). At the present stage of knowledge, about primary personality factors, it seems quite unlikely that they will show such equality. Consequently, we shall undertake in the following section to generalize the coefficient of profile similarity,  $r_p$ , to meet the need for unequal weighting, as well as to add other needed flexibilities.

### 5. Varieties Within the Family of Profile Similarity Coefficients

From the preceding section, it becomes evident that the commonly used profile similarity coefficient,  $r_p$ , is really one of a special case: one of many possible formulae within a family of coefficients. There could, for example, be weights and polynomial expressions for calculating similarity (or "distance") with respect to all kinds of relations to criteria and particular combinations of criteria. The ordinary  $r_p$  is a special case from these in operating with a linear combination of squared differences which gives equal weight to all dimensions. Furthermore, its quadratic form specifically assumes a non-linear, parabolic relation of individual traits to criteria. This means that in evaluating the extent to which an individual belongs to a clinical syndrome "type" (or to take another example, his



"adjustment" to the ideal, stable profile of those in a given occupation), it (a) penalizes equally for under and overshooting the criterion value and (b) does so in terms of the square of the deviation involved.

Obviously, modifications could be made in the similarity formula to fit all kinds of assumptions about the relation of trait to performance, which could be expressed in various polynomials relating profile factor scores to the criterion. Indeed, one instance of modifying the  $r_p$  index which may be briefly mentioned, because it is actually more consistent with the widely psychologically used linear factor specification equation than is  $r_p$ , is what we shall distinguish as the coefficient of linear similarity,  $r_s$ . In this, the signs of the d's, standard score differences of two persons on the succession of factors, are preserved in the addition and the coefficient will indicate not only the degree of similarity of two persons, but also which is positive (higher) relative to the other. It is defined by where there

$$r_{sp_1p_2} = \frac{2k \sum b - \sum bd}{2k \sum b + \sum bd} \quad (2)$$

are  $k$  profile elements, the  $b$ 's are the factor weights and the  $d$ 's the differences on the factors, person  $p$  always being subtracted from person  $p_2$ .

The expression,  $r_s$ , will preserve consistency with the familiar linear specification equation, but the similarities thus calculated will lose any relation to an Euclidean distance. Yet another member of the profile similarity family, and one which succeeds in approaching Euclidean distance properties even better than  $r_p$  (see Figure 2), is what we may call the coefficient of nearness,  $r_n$ , defined as follows:<sup>4</sup>

$$r_{np_1p_2} = \frac{\sqrt{2k} - \sqrt{\sum d^2}}{\sqrt{2k} + \sqrt{\sum d^2}} \quad (3)$$

---

<sup>4</sup>Strictly the expected value of  $\sqrt{\sum d^2}$  is  $\sqrt{2k} (1 - 1/4k + 1/32k^2 + \frac{1}{128}k^3 + \dots)$ ; but  $\sqrt{2k}$  is a close enough approximation if  $k$  is not too small.

---

The greater conformity of  $r_n$  to Euclidean space (i.e., to a generalized D) is shown in the graphs of Figure 2. Like all members of the  $r_p$  family,  $r_n$  has a numerically immediate meaning as a similarity coefficient in that it yields 0 when the relationship is an average, random value; it reaches +1.0 for exact likeness; and approaches -1.0 for maximum dissimilarity.

What recommends it less than  $r_p$  is that its distribution skews more, approaching -1.0 very slowly. At a 5 sigma difference on every element it is still only approximately -0.6. This slowness to approach -1.0 may express a necessary truth namely that in any ordinary biological or social population extreme opposites are much more rare than individuals who closely resemble each others. This inference, as well as certain other properties of  $r_n$  and  $r_p$  warn us that in averaging and in other manipulations of pattern similarity coefficients we need to watch certain pitfalls. Since there has been practically no reported experience with  $r_n$ , whereas  $r_p$  has been appreciably tried out, our further discussion will keep to the latter, considering the further issues of weighting and obliquity only in regard to the generalized  $r_p$  formula.

Published uses of  $r_p$  to date have used the specific, non-generalized form, which has two main assumptions: (1) that the factor measurements are orthogonal, and (2) that the elements (factors) are to be given equal weight. Yet most known personality and ability source traits stand obliquely to one another so that assigning nominal weights to items would not give equal statistical weights. And often we wish to give them known unequal weights, which, incidentally, implies also that we are giving certain weights to the higher strata (Cattell, 1965) factors arising from the oblique factors.

Probably it would be correct to say that most psychologists implicitly assume in comparing personalities that they want to give equal weight to each and every behaviour in real life, i.e., to consider the realm of criterion performance as the basis for perspective. If so, they should recognize that to achieve this goal it will nevertheless be necessary to give unequal weights to the factors. Unequal weights are necessary because in predicting variables constituting a stratified sampling of the universe of behaviour we are likely to find some factors more "important" than others. A precise expression (granted an available defined total population of variables) for the differing importance of individual factors can be obtained from estimates of the mean variance contribution of each factor across the population of variables, i.e., by the root average squared sums of the factor loadings for the given factor (the "latent root" in the orthogonal case) as follows:

$$w_{fj} = \sqrt{\frac{\sum_{x=1}^n b_{jx}^2}{n}} \quad (4)$$

where  $b_{jx}$  is the loading of variable  $x$  on factor  $j$ .

Other rationales for weighting may be proposed, but regardless of their nature we shall need a generalized  $r_p$  for any obliquity and any weight. Let us begin with the essential form behind equation (1), namely,

$$r_{pxy} = \frac{E_k - d_{xy}^2}{E_k - d_{xy}^2} \quad (5)$$

where  $d^2_{xy}$  is the squared distance apart of two people,  $x$  and  $y$ , in a  $k$  dimensional Euclidean space and  $E_k$  is the expected distance for  $k$  dimensions. But  $d^2_{xy}$  can no longer be simply  $\sum_{j=1}^k (x_j - y_j)^2$  (or, in matrix notation,

$z'_d(xy) z_d(xy)$ ). For we must now take into account the correlations,  $r_{f_j f_l}$  between the source traits (factors)  $j$  and  $l$ , and others, which we may write as the usual matrix  $R_f$ , and we must also include the weights assigned to the factors, which we will write into the  $k$  by  $k$  diagonal matrix  $D_w$ . Then:

$$d^2(xy) = z'_d(xy) D_w^2 R_f D_w^2 z_d(xy) \quad (6)$$

The expected value of  $d^2(xy)$  is no longer  $2k$ , but is:

$$E_k = \text{trace} \left( D_z^{-1} L' D_w^2 R_f D_w^2 L D_z^{-1} \right) \quad (7)$$

where  $D$  is the diagonal matrix of latent roots of  $z'_d(xy) z_d(xy)$  and  $L$  the matrix of the associated latent roots.

If one wishes to revert to the special case so far employed -- the orthogonal, equal weight  $r_p$  -- it is easily done by inserting  $r = 0$  and  $w = 1$  in the above. The computing convenience of the orthogonal approximation we have been using (acceptable when only minor obliquities exist) is thus very substantial and attractive; for the user of the oblique formula is compelled to work out afresh for each case the complex expression  $z'_d(xy) D_w^2 R_f D_w^2 z_d(xy)$ . To employ the simple (orthogonal) approximation, on the other hand, it suffices only to enter a nomograph with the individual  $d^2$  value (Table in Cattell and Eber, 1966). However, with the help of a computer program, based on (7), the use of the exact oblique formula, even with quite large numbers of individual cases, presents no real problem.

The formula for the profile nearness coefficient -- (3) above, using  $d$ 's without signs, instead of  $d^2$  -- when correspondingly adapted to specific source trait obliquities and weightings becomes:

$$E_k = \frac{(z'_d(xy) D_w^2 R_f D_w^2 z_d(xy))^{\frac{1}{2}}}{(z'_d(xy) D_w^2 R_f D_w^2 z_d(xy))^{\frac{1}{2}}} \quad (8)$$

Here, to a first approximation:

$$E_k = (\text{trace} (D_z^{-1} L' D_w^2 R_f D_w^2 L D_z^{-1}))^{\frac{1}{2}} \quad (9)$$

The distribution and significance limits for  $r_n$ , corresponding to those obtained for  $r_p$  (Horn, 1961) remain to be worked out, so the further steps and applications we now propose to follow are best considered to employ  $r_p$ .

## 6. THE OPERATIONAL DEFINITION OF PHENOMENAL AND NUCLEAR CLUSTERS (OR CLIQUES) IN AN INCIDENCE MATRIX

With the above treatment of the problem of calculating similarity (a reciprocal function of distance) as such, for any two ids, we are ready for the operations in finding types. In the first step from this similarity value, be it  $r_p$ ,  $r_n$ ,  $D$  or any other consistent concept -- toward classifying people in types one must introduce a limiting value -- arbitrary or natural -- in order to shift from a quantitative or parametric to a qualitative or categorical treatment. At some point one must end by speaking of people as "in" or "out" of a type, though degrees of belonging may also be used later.

Although we must never lose sight of the metric origin of the cutting point, and the way in which its choice can affect the grouping, yet we now propose to convert the  $Q$  matrix of  $r_p$ 's into an "incidence matrix". Therein, if a certain limiting positive  $r_p$  value is exceeded in the original  $Q$  matrix a unity is entered, to designate a linkage, whereas if  $r_p$  is not positive, or is below this significance a zero is entered in the cell to show that the two people are unrelated. There will thus be no negative values, but only 0's and positive unities in the incidence matrix.<sup>5</sup>

---

<sup>5</sup>This is perhaps the place to point out that the reciprocity of  $R$ - and  $Q$ -technique practices breaks down in one important respect: one can meaningfully reflect tests but not people. Consequently, one cannot meaningfully reflect  $r_p$  coefficients signs (to make them positive) by reflecting one of the two people. It is true that conceptually we may do so, and that we recognize a special logical affinity of opposites, as when we talk in one breath of angels and devils, and theology insists that Lucifer had to be a fallen angel. But what is the opposite of a chair? Opposites to existing objects may be mathematically conceivable, by logical fiat, but not consistent or conceivable in scientific properties. Certainly for most objects opposites simply do not exist in any actual world of data. So, like D'Artegnon, we may assert "Le diable est mort" without becoming atheists! In short, in the whole process of mapping similars we are not required to consider opposites, and certainly we are not permitted to make reflections in  $Q$ -matrix id entries. Parenthetically, with correlations of persons, reflecting even a test upsets the inter-person similarity value, as pointed out by Cattell (1952a) in the early discussions of  $Q$ -technique, and illustrated pointedly in a recent paper by Howard and Diesenhuis (1965).

---

Once the abstraction of the incidence matrix is reached, with "links" taking the place of similarity values, both the scientific model and the computer program we are developing for it take on broader reference and utility. In most respects they apply both to the personality and cultural psychologist's (as well as the biologist's) need to find types and to the sociologists need for an objective basis for locating cliques and communication networks (Cattell, 1963). These aims formally express themselves in finding what we have called stats (not segregates). Within stats themselves,

however, two distinct sub-concepts are now needed: phenomenal clusters (or stats) and nuclear clusters (or stats). A phenomenal cluster (henceforth p-cluster for short) corresponds to what is perhaps the simplest operational definition of a homostat as a homogeneous set of ids. It is defined as a set of ids each of which is linked to every other (and which does not exclude any other id similarly linked to the set). Spatially this means that all fall within a hypersphere of diameter fixed by the similarity coefficient level accepted as a link. The word "phenomenal" is used because such a cluster is directly obvious and given in the data relations, whereas a nuclear cluster (henceforth n-cluster) as we shall see in a moment, has a less direct definition, because it requires an extra operation of abstraction.

Obviously the number and the nature of the p-clusters found in given data will alter with the id from which search is started and with the cutting point on  $r_p$  which is used as a similarity limit, i.e., translates as a link in the incidence matrix. Different groupings will appear as the limit is dropped, just as the sand bars in an estuary change shape with the tide. Some typologists both in psychology and biology, have been frankly arbitrary, setting some value from +0.5 to +0.8 as a limit according to "judgement". Since arbitrariness of this degree is unsatisfactory, two possibilities of objectivity need to be considered. First, one may shift the decision to a decision on the number of types one expects to find, which is the inverse of the average size of a type, in terms of percentage of the total population included. (If one visualizes a two-space filled with adjoining circles, now large, now small, he will see what the alternatives mean.) This remains on a completely arbitrary basis, but it is one which can be referred more directly to the goals of systematics in the given field than can the  $r_p$  value per se. Secondly, one can take a cutting point dictated by the distribution of the distances in the ids themselves in the sample. For example, in a sample of 100 a critical distance might be chosen such that most ids will stand as isolates. (Or, in general, most clusters will contain only 1 per cent of the population.) This recognizes the relativity of types, e.g., that a hundred people shoulder to shoulder counts as a crowd in Times Square or Picadilly, but six people within sight of one another indicates a group if found in the Sahara. In the last resort this encounters the same arbitrary decision as the first method: "What fraction of your population do you want to include in types?" However, it does suggest an initial objective operation, namely, to take as the cut off point the mean of the positive  $r_p$ 's in the matrix, or to take the mean of the  $r_p$ 's from random normal deviates for  $k$  profile elements. This latter, incidentally, will not be exactly zero, but it will make roughly half the links significant. Table 1 shows values thus generated, to illustrate their dependence on the number of elements.

(Insert Table 1 here)

Table 1 answers the question sometimes raised: "If we take  $n$  times as many people randomly distributed in the same space will not the average distance of each person from every other be correspondingly reduced?" Here Mahalanobis'  $D$  will be more susceptible to sampling, but  $r_p$  scarcely at all, as Table 1 shows, for although there will be an increase in the total number of similar people there will be a corresponding increase of those who are dissimilar, i.e., mutually correlating 0 to -1.0. However, for a given

limit of admission by  $r_p$  to a homostat more people will, of course, be included in absolute terms, if the population structure remains the same, with a large than a small sample. Sampling laws for stats and sits remain to be worked out, but to a close approximation multiplying the sample size by  $n$  will multiply the number in any given diameter of stat by  $n$ . Consequently all type structure statements should at some stage be converted to percentages and further analysis pursued on that basis.

Granted an agreed critical cutting point on  $r_p$ , leading to a linkage Q-matrix, by what systematic operations can one derive the p-clusters? A Boolean algorithm for this purpose will be described in the next section, but here we have still to complete the conceptual distinction of phenomenal and nuclear clusters and so for the moment we shall take a small example in Figure 3 in which the phenomenal cluster is obvious from Table 2. In fact three instances of p-clusters are illustrated topologically in Figure 3, namely, a b e f g; a b c d h; a b c d e i.

(Insert Figure 3 here)

It will be noticed, however, that the first two p-clusters overlap with respect to ids a and b. That is to say, a and b are linked in all necessary ways for a p-cluster with e, f and g on the one hand and c, d and h on the other; but c, d and h are not linked with f, g and e. The term nuclear clusters, or n-cluster is therefore given to a, b. If one now considers the third p-cluster (No. 2) in Figure 3(i), he will note that the nuclear cluster concept can get complicated, to the extent that "orders" of nuclear clusters must be introduced, according to the number of p-cluster overlaps involved. Thus c and d are in a two p-cluster n-cluster, but a and b are sustained by a three p-cluster overlap. An n-cluster finishes by being more than the definition of a simple stat: it is a stat with additional "structural" properties.

(Insert Table 2 here)

As instances (i) and (ii) in Figure 3 suggest, the structural varieties of n-clusters according to the associated form of relation of p-clusters can be very diverse. And since the description of a population sample in terms of p-clusters alone may vary (as pointed out, by our tides and sandbanks analogy, showing groupings to alter according to the cut off level on  $r_p$ ), the n-cluster description will also change with the critical cut-off value. Consequently, to approach an adequate description of a domain it is desirable to present groupings at each of several, say, three standard levels (for which experience suggests  $r_p = 0.2, 0.5$  and  $0.8$ ), as a cartographer presents contour lines only at standard levels. For convenience these levels may need adjusting to the parametric properties of the given data as in our analogy of the Sahara and Times Square. On the other hand, if certain standard  $r_p$  levels could be agreed upon in type research generally, it would advantageously permit comparisons of various domains for what in our introduction we briefly called texture. Texture can now be given more specifically the meaning of the number of p-clusters, of various percentage sizes, at various cutting levels, plus the n-cluster sizes at various numbers of p-cluster

overlaps, etc. With this glance at the number of summarizing statements required, our introductory statement about the inappropriateness of hoping for a simple, single, boiled-down mathematical statement when mapping clouds in  $k$ -dimensions will be more self evident: we are dealing with topography.

## 7. THE BOOLEAN CLUSTER SEARCH ALGORITHM FOR FINDING STATS

Let us now consider the logical and computational requirements in proceeding from a given incidence matrix, as in Table 2, to a statement about stats (as  $n$ - and  $p$ - clusters) such as is summarized visually in Figure 2. It is this step which will provide the basis of the Taxonome computer program. For finding clusters in correlation matrices, Cattell (1952) originally proposed the ramifying linkage method algorithm, but subsequent use showed the need for an additional step, and we now call the revised method the Boolean cluster search method.

It still begins with the ramifying linkage method which proceeds from the original  $Q$  matrix (henceforth  $Q_0$  for the basic matrix, to distinguish it from subsequent derivatives, analogously to  $V_0, V_1, V_{II}$ , etc., in factor analysis). Herein one works sequentially through the given links for one person after another, i.e., column by column or row by row in Table 2, at each step deleting any ids not directly linked to those found in the earlier columns. It will be found that in this comparatively simple example the ramifying linkage method alone leads reliably to the clusters shown in Figure 2. However, for the sake of illustrating certain higher derivatives we shall turn to a new but still small example presented by Table 3, to illustrate the need for the full Boolean process. Beginning with the incidence matrix among ten ids in  $Q_0$ , the process (and the subsequent computer program first scans column 1 and thus notes the set of persons related to person 1, namely, persons 5 and 7. It proceeds next to column 5 and notes that person 5 is related to person 7; so 1, 5 and 7 form a cluster.

(Insert Table 3 here)

Incidentally, in setting up the  $Q_0$  matrix a triangular form is sufficient, for if ids  $i$  and  $j$  are related, then the  $(i,j)$  and the  $(j,i)$  elements of  $Q_0$  are 1, but computationally it is more convenient to use the whole matrix, recognizing, however, that this may result in our finding the same cluster twice.

From  $Q_0$  our aim is to produce a matrix  $G_1$  (for "grouping matrix") giving an initial statement of existing clusters according to the ramifying linkage method. As we encounter each link in column 1 we must decide if the id (person) concerned is also linked directly with other persons having links in that column. To decide this we must see if for every entry of unity above him there exists a corresponding entry of unity in his row (or equivalently, column) of  $Q_0$ . (The method as originally described by Cattell required comparison with all unit entries below the one being considered, a logically equivalent procedure though slightly less efficient for computing.) So, in  $Q_0$  of Table 3, we see that 5 is linked to 1, then



going down the column to the next unit entry we find that 7 also belongs to the group since when we look along the 7 row there is one unity in column 5, i.e., a link of the two persons already included in the group. Column 1, for a contingent group, is therefore started in matrix  $G_1$ .

Going next to column 2 of  $Q_0$  we find persons 2 and 6 form a group, from column 3 that 3 and 7 form a group, and these are entered in  $G_1$  as columns II and III. Column 4 contains a single unity and need not be considered. Working down column 5 we include 1 and 5, but on examining person 6 we find a zero in the first column of row 6, so 6 does not belong in the group and the unity corresponding to person 6 in  $G_1$  is changed to zero. 7 is related to 1 and 5, and so is included. However, the group now found is identical to group 1 and so we do not include it in  $G_1$ . Similarly, we work through columns 6 to 10, finding in all the five distinct groups listed in Table 3 as the columns of  $G_1$ .<sup>6</sup>

---

6. Two points must be noted about the ramifying linkage method. Firstly, some of the clusters initially found may be subsets of other clusters. This presents no problem. Secondly, due to the sequential nature of the procedure, not all clusters may initially be found, at least where certain unusual configurations exist. (This is the reason for the next step from the  $G_1$  matrix.) Thus in Table 3(a) the group consisting of persons 5, 6 and 7 is not found. We do not include phenomenal clusters of only one person, which correspond to a column with only a diagonal element that is non-zero, e.g., columns 4 and 8 of Table 3,  $Q_0$ .

---

Actually, the ramifying linkage method is best regarded as a first step, in the way that taking out a first factor came to be regarded as only the first step in a multiple factor analysis. Indeed, the formal similarity to factor analytic steps is appreciable, for our procedure is to set down a first phenomenal cluster matrix,  $G_1$ , from the ramifying linkage "extraction" process, and make therefrom a product matrix,  $Q_1$ , which, subtracted from  $Q_0$ , leaves a first residual,  $Q_2$ . Thus, step 2 in Table 3 is:

$$Q_1 = G_1 \cdot G'_1, \quad (10)$$

where the prime denotes a transpose and the period denotes Boolean matrix multiplication, i.e., a matrix multiplication with arithmetic addition and multiplication replaced by logical addition ('or') and multiplication ('and').

If  $G_1$  should contain all p-clusters, then we must have

$$Q_0 = Q_1$$

since a link (other than a diagonal one) in  $Q_0$  indicates that two persons are related, and so they must appear together in at least one phenomenal



cluster. The operation  $G_1 \cdot G'_1$  simply determines which persons appear together in phenomenal clusters. Table 3(2) gives  $Q_1$  for the example. Zero's in  $Q_1$  corresponding to unities in  $Q_0$  have been denoted by x's, indicating that in this case not all phenomenal clusters have been found. Now the new "residual" incidence matrix,  $Q_2$ , is formed from the x's of  $Q_1$ , plus any element in their columns (a) that was unity in  $Q_0$ , and (b) for which there is also an x in its row of  $Q_1$ . Such an element might form a phenomenal cluster with the x's and so needs to be included. Table 3(3) gives the  $Q_2$  for the example. Using the ramifying linkage method we now find additional phenomenal clusters -- in this case one, No. VI -- which we include with those already found to form  $G_2$ .

Then,

$$Q_2 = G_2 \cdot G'_2 \quad (11)$$

and

$$Q_2 = Q_1$$

if all phenomenal clusters have been found. We proceed in this way until we find a  $G_n$  such that  $Q_n = Q_{n-1}$ , except possibly for some diagonal elements. In the example, Table 3,  $Q_3 = Q_2$  except for the (4,4) and (8,8) elements, so  $G_2$  contains all the phenomenal clusters in  $Q_0$ .

## 8. PROCEEDING FROM STATS TO AITS,

### TO DENDROGRAMS AND TO TEXTURES

By adding a simple search and counting procedure which will list the overlaps among the p-clusters for the algorithm just described, the findings up to this stage can be systematically recorded, as briefly indicated above. They will finally appear as a print-out of (a) p-clusters and (b) n-clusters. To be comprehensive of possibly needed information these lists will in detail comprise:

(1) For p-clusters: (i) a listing of actual id members, (ii) arranged in order of size from 2 membership upward, (iii) attachment of identifying numbers to clusters, and (iv) expression of size in percentages of same and calculation of the distribution by cluster frequency, as shown in Table 4.

(Insert Table 4 here)

(2) For n-clusters: (i) a listing of actual id members, (ii) attachment of identifying numbers to cluster, (iii) arrangement in this case in a two-way table, by size (expressed as percentage) and by number of p-clusters involved in the overlap, (iv) a distribution analysis on both of these. For the data of Table 2 this is shown in Table 5.

(Insert Table 5 here)

To complete the general statement at the stat level, these two tables must be repeated for whatever number of cutting levels on  $r_p$  one feels to be necessary, probably three as indicated above.

The investigator will want to know how far he can make inferences from this sample result (our present taxonome program handles 140 cases) to the population. It should be noted that stats are subject to the particularity of the sample in two senses, first the ordinary sampling sense and secondly by the dependence of the center and boundaries of the state upon the id with which one begins. As to the former, since no theoretical mathematical statistical treatment is yet available investigators had best develop estimates of standard errors for sampling by Monte Carlo methods. As to the latter, which will become clearer as we discuss sits, the problem arises from the fact that the center and boundary of a stat depends upon the id with which we happen to start the process.

The final list of stats will escape any bias from this source on the alternative "cartet" procedure, and it will do so in the id-similarity procedure here too, because all possible commencement points have been included. But it does so at the cost of generating a possibly bewildering number of overlapping stats in the p-cluster list above. For the number of p-clusters, namely  $\binom{n}{x}$  where  $x$  is the number encircled at the given distance diameter, could decidedly exceed the original number of ids! Tables 4 and 5 are for a small example; with one of moderate size the investigator may well ask whether the procedure was intended to produce data reduction!

To use the stat lists the investigator will need to look at the distribution and ask what fraction of the population he wants in types. He must also remember that a large cluster really means a dense cluster, since all p-cluster diameters are the same. Possibly he will want to use the non-overlapping highest density clusters which cover at least 60 per cent of the population. Or again he may want n-clusters simultaneously above a certain density (size) and a certain p-cluster overlap frequency. For example, by rejecting from List 1 (Table 4) all p-clusters with fewer ids than are shown by the two or three largest orders, one would get just two types (dotted circles) in A, Figure 4, and two or three at the heart of B. The decision must depend upon texture, and here texture begins to assume a definable meaning. It resides in the evidence of the p- and n-clusters in the stat list (Tables 4 and 5) as to how people are distributed between small and large clusters, how much overlap occurs respectively with small and large, and whether any hierarchical, dendritic structure is apparent.

Let us now turn to locating sits (segregates). We are bound to begin with stats, yet utilizing this information is like seeking to locate the objects in a large picture in a darkened room with a flashlight throwing only a small circle of light. The circles -- the p-cluster stats -- will pick up the object only piecemeal and a method will be necessary to put the pieces together.

Consider a simplified case as in Figure 4, with people spaced as shown, yielding two dense segregates A and B on an otherwise "dilute" field.

Let us assume search is made with three levels of  $r_p$  cut off, namely, +0.8, +0.5 and +0.3 corresponding, in two-dimensional space, approximately to circles of the sizes shown. The first will give practically no p-clusters in the field, since only in the A and B clumps will it span two cases. The lowest cutting point (+0.3), on the other hand will bring every one of the ids into one cluster or another, as illustrated by the span of its circle at the top left. If we followed through with this, as we have with the middle value circles (0.5), the whole space would be covered with circles representing p-clusters, though the n-clusters would only appear where the A and B segregates stand.

(Insert Figure 4 here)

At this point the question might be raised whether an n-cluster is not conceptually equivalent to an ait, but the answer must be no. For if an n-cluster is confined to what is common to p-clusters of a certain size it cannot itself exceed that size -- and an extended ait will commonly need to do so. Nevertheless, and incidentally, one sees many instances in the literature where investigators have adopted stat search procedures despite their conceptualization of their problem clearly indicating that they are looking for aits. It will help to clarify this point to observe that in Figure 4 the aits are the masses A and B. In this case it happens that by confining oneself to the larger state, i.e., those at the top of Tables 4 and 5, one finds in this case the heart of these two segregated masses. But it will not always be so, as a glance at a chain, as in Figure 3(ii) will remind us. There the nuclear clusters are not central. It must also be remembered that a larger number of people collected in a stat by the above operations is not an indication that it is large (in the sense of covering large areas of behavior) but only that people are very dense in the given region -- which is possibly quite small. Always it must be borne in mind that in a very extended ait the last members may have negligible, zero or even negative resemblance to the first. For example, it might be said of a certain religious group X that it has a tremendous range of values and practices, so that despite continuity and coherence in the chain of resemblance of members an extreme X may be more like a member of another religion, Y, than like members at the other wing of his own religion. This statement is illustrated by B<sub>3</sub> and C<sub>1</sub> members being, in Figure 1, in the same stat, No. 2, but in different aits. Despite this lack of homogeneity present in the stat the recognition of aits is important in many aspects of social, educational and clinical psychology.

The operation we have devised for objectively locating segregates consists of first finding stats and then setting up a stat contiguity matrix, very similar to the Q matrix of linkage among ids, except that it now represents linkage (interpreted as a sufficient degree of overlap) among p-clusters. Before this  $Q_c$  (relations among clusters) matrix can be set up, one must settle, from the evidence on the general texture of the domain given by the equivalents of Tables 4 and 5 above, on: (1) the cutting limit of  $r_p$ ; (2) the densities (numbers of ids in a stat) to be accepted (clusters of only 2 and 3 persons would normally be rejected as too unstable); and (3) the amount of overlap to be accepted as evidence

of linkage (one id might be too subject to sampling variation; an overlap of 2, 3 or more seems more appropriate).

The taxonome program as now set up accepts its instructions on these limits from values inserted for the particular problem by the experimenter and then presents a  $Q_c$  incidence matrix among p-clusters. But from that point on, the search made in  $Q_c$  is quite different from the Boolean Cluster Search Algorithm used in  $Q_0$  for finding stats. Now we are no longer interested in maintaining the condition that every member (a member now being itself a cluster) shall be linked with every other. Instead we are interested in segregating all the ids (clusters) which are continuously connected with one another through any intermediate ids maintaining the stipulated degree of resemblance. The procedure now requires that we go down a column of  $Q_c$ , find the other ids (in this case clusters) linked to it, and then pursue all its connections, and so on for further additions to the family. Thus even the shape of an octopus would be recognized by this procedure, provided the tentacles at no point get so thin as to preclude visible overlap -- by the stat size which means "visibility". This we may call continuous connectedness analysis. The further issues of texture tactics and boundaries presented by such problems as this last will be discussed in a moment, but first the main "Segregate Search" procedure will be described.

(Insert Table 6 here)

Again the program employs Boolean algebra concepts. The investigator (or, in our program, the computer) proceeds systematically from column 1 down the other columns of an incidence matrix,  $Q_c$ . This is derived from the data of the earlier (individual person) example, summarized in Tables 4 and 5, via a pre-incidence matrix, (a), in Table 6, which gives the numbers involved in the cluster overlaps. Proceeding down the first column of the  $Q_c$  matrix one accrues the ids in the rows corresponding to the incidence signs. At each such id one runs across the row and accumulates new columns where incidence signs occur, following these likewise across rows which are not null. Thus in Table 6(b), columns 1, 2 and 3 begin to form a segregate but the intersection of this with 4, 5 and 6 is null. Starting again with 4 one finishes with 4, 5 and 6. Illustrated in Boolean terms, if the columns were as in (a) the Boolean product would be zero, and we should proceed no further. In (b) on the other hand, it is not null, so we proceed to Boolean addition to form the new segregate, shown in the last column of Table 7.

(Insert Table 7 here)

Obviously, the detail in the picture of segregates will, as in a photograph depend on the size of the grain. A glance at Figure 5 will show that if the smallest circle ( $r_p = 0.8$ ) were used the isthmus between the two parts of the dumbbell shaped A segregate would not appear; though, on the other hand, a gain would result from certain fringe persons around A and B being dropped who perhaps could be said not really to belong.

It may be asked why the search for aits is not carried out by applying what we have called the continuous connectedness analysis (Table 7) directly

to resemblances of individuals in the manner that the Boolean Cluster Search has been used for matrix  $Q_0$  (Table 3), or Table 2. Our answer is that a single individual is altogether too slender a datum, in view of sampling error, upon which to rest connectedness. Thus, at the cluster search stage the elimination of smaller, e.g., two-man, clusters from List 1 (Table 4) is likely to take care of sampling error "artefacts" in the original data, whereas it would be difficult to eliminate one man threads in the continuous connectedness analysis. Accordingly it has seemed better to locate stats and then use these as units in recognizing the continuous connectedness sought in sits. However, more could be said, and certainly more needs to be done in the way of experiment upon the effects of adjusting the size of stat diameter to the texture of the domain, when seeking sits.

## 9. TRIAL OF TAXONOME ON REAL DATA AND PLASMODES

A description of the technical flow chart of the computer program built by us on the above principles is set out elsewhere (Cattell and Coulter. This journal. p. ). It is to be hoped that others, in experimenting with its use, will develop ways of finding the best parameters in the program suitable for various textures and kinds of data. Here we report only on two sufficiently diverse practical examples to show that Taxonome works to a reasonable degree. A trial of the algorithm, but by desk computer, was made by Cattell (1950) soon after devising  $r_p$ , on an example of general interest, namely, the classifying of national culture patterns into types of "civilizations," to check on Toynbee's speculations. Using a twelve factor profile for each of 69 countries Cattell obtained some ten phenomenal clusters centering on two nuclear clusters. Four of the former are set out in Table 8 for illustration.

(Insert Table 8 here)

It will be seen that these blindly statistically obtained stats make sense in terms of the usual socio-historico-anthropological evaluations. Thus encouraged, we proceeded (albeit with too many interruptions) to the present taxonome, which is now being tried by us on a number of plasmodes. (Plasmodes have been defined (Cattell, 1966) as arrangements of specific numerical values to fit a mathematico-theoretical model. They are useful for gaining new insights into the working of a model and for trying out computer programs intended to analyze data according to such a model.)

While waiting to complete studies on strategically chosen plasmodes we decided to try a nursery model, using as data 29 vessels from "Jane's Fighting Ships" (1964-65) representing four distinct types of craft -- aircraft carriers (5), destroyers (4), submarines (10) and frigates (10). Twelve measures were used in the profile of each, for the  $r_p$  calculations: (1) displacement; (2) length; (3) beam; armament in number of, (4) light, (5) medium, (6) heavy and (7) very heavy guns, (8) the complement, (9) maximum speed, (10) submersibility, (11) continuity of deck construction, and (12) whether no, some or many aircraft were carried.

(Insert Table 9 here)

The incidence matrix (Table 9(a)) suggests to the eye that the breakdown into four classes will be reasonably good, but the actual p-cluster output (Table 9(b)) indicates 9 clusters. Three of these are clearly the destroyers, submarines and frigates, but the aircraft carriers have broken into 3 p-clusters which, later, however, yield a single nuclear cluster.

Further, more complete applications, which cannot be described in this introductory paper, are being reported elsewhere.

## 10. SUMMARY

(1) The most useful general concept of a type requires that it be defined as the central profile in a high, "modal" frequency (unusually high density) of individuals in a multi-dimensional distribution.

(2) Two sub-concepts can be operationally distinguished within the notion of type so defined: (a) the stat (for homostat) -- a homogeneous group in which each member stands at less than a given distance (the same for all) from all other members, and (b) the sit (for segregate) -- a continuous but not homogeneous group in which each member is nearer to at least one other member than he is to ids outside the group.

(3) State (homostats) and sits (segregates) can be found by either "inter-id relation" or "density in space" (cartet count) methods, the former being pursued here. This requires a measure of similarity (the opposite of distance apart in the given space) for every pair of ids (i.e., persons, groups, processes, etc.). Reasons are given for preferring as a similarity index the family of profile similarity coefficients ( $r_p$ ,  $r_n$ ,  $r_s$ , etc.) to the correlation coefficient, Mahalanobis' D, or other coefficients sometimes proposed for this purpose.

(4) Similarity can be considered either in regard to (a) some specific criterion performance or averaged group of performances. This leads to classification of ids by their effects or works, or to (b) general purpose dimensions, resting on the concept of sampling a personality sphere or a population of variables. This implies classification according to the "thing in itself".

(5) In the last resort these need the same mathematical treatment, since even the "thing in itself" concept implies some weighting in the personality sphere. Formulae are presented for inter-id similarity indices based on the principal useful alternative assumptions, e.g., regarding linear and parabolic relations to criteria, and generalizing the original profile similarity coefficient  $r_p$  to any correlations among profile elements and any weights.

(6) The discovery of stats begins with a Q-matrix of  $r_p$ 's among ids. At each of two or three cutting points for  $r_p$  this is converted to an incidence matrix. A Boolean algorithm, based on what was called the "ramifying linkage method", objectively sorts the data into phenomenal clusters. An operational distinction has to be made between phenomenal (p-) clusters and nuclear (n-) clusters which have quite different properties. The conclusion of the search for stats consists of one list of phenomenal clusters, by size and specific members, and one list of nuclear clusters, by size, number of overlapping clusters involved, and specific members. These lists, which give the "texture" of the domain, can be voluminous and require that the investigator select an importance level to reduce the number of concepts to be handled.

(7) The discovery of sits (segregates) begins with a  $Q_c$  matrix of overlap among phenomenal clusters which is converted to an incidence "contiguity" matrix and operated upon by a Boolean analysis for continuous connections.

Experiment is needed to find the best rules for size of stats to be used in seeking aits.

(8) The concepts and principles of analysis have been incorporated in a computer program (for the IBM 7094 initially) which has been shown to work on two concrete examples, though experiment on others, adjusting the parameters optimally, especially to minimize sampling effects, remains to be done. Unless a theoretical mathematical-statistical solution is soon found, Monte Carlo methods should be employed to establish sample inference limits in this field.

(9) Over and above the finding of particular stats and aits a search for types the taxonome method aims to describe the texture of a domain. We have referred to texture by the analogy of the meteorologist's use of cumulus, alto-stratus, etc., to describe cloud formations. Segregates can appear as small or large, even or unevenly spaced, massed or in chains, etc. Operationally, texture will broadly be defined by comparisons of structure at different cutting levels, by the ratio of nuclear to phenomenal clusters, by the degree of compactness<sup>7</sup> of aits, and by the amount of hierarchical structure

---

<sup>7</sup> An index of compactness can be obtained by dividing the total number of ties (incidence matrix) involved in a segregate by the total number possible --  $n_c2$ , where  $n$  is the number of ids involved in the segregate.

---

discernible among them, as in the biologists' dendrograms. The ascertaining of the last has not been described in detail, but clearly involves a "second-stratum" repetition of the type search carried out upon the patterns representing the central tendencies in the type groupings first found.

(10) The empirical search for types will naturally need to proceed hand in hand with inductive and deductive theory development on the origins, interactions and natural history of types. A theory of three sources of type structures is stated and one of them suggests that the use of type concepts in psychology is likely to become tied to the development of non-linear specification equations.

The writers gratefully acknowledge that this investigation was supported in part by Public Health Service Research Grant No. MH 1733-09. They are indebted also to Professor Peter Schoenemann for help and advice on the early stages of the program.



## References

- Burt, C. L. (1937). Correlations between persons. *Brit. J. Psychol.*, 28, 59-96.
- Burt, C. L. (1940). *Factors of the Mind*. London: Univ. of London Press.
- Cattell, R. B. (1946). *The Description and Measurement of Personality*. New York: World Book.
- Cattell, R. B. (1949).  $r_p$  and other coefficients of pattern similarity. *Psychometrika*, 14, 279-298.
- Cattell, R. B. (1950). The principal culture patterns discoverable in the syntal dimensions of existing nations. *J. soc. Psychol.*, 32, 215-253.
- Cattell, R. B. (1951). On the disuse and misuse of P, Q,  $Q_8$ , and O techniques in clinical psychology. *J. clin. Psychol.*, 7, 203-214.
- Cattell, R. B. (1952a). The three basic factor-analytic research designs -- their interrelations and derivatives. *Psychol. Bull.*, 49, 499-520.
- Cattell, R. B. (1952b). *Factor Analysis*. New York: Harper.
- Cattell, R. B. (1957). *Personality and Motivation Structure and Measurement*. New York: World Book.
- Cattell, R. B. (1965). Higher order factor structures and reticular-vs-hierarchical formulae for their interpretation. In C. Banks and P. L. Broadhurst, *Studies in Psychology Presented to Cyril Burt*. London: Univ. of London Press, pp. 223-266.
- Cattell, R. B. (1966). (ed.) *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally.
- Cattell, R. B. and Coulter, M. A. (In preparation). Nomographs, significance tables, etc., for use of the profile similarity and nearness coefficients,  $r_p$  and  $r_n$ .
- Cattell, R. B., and Eber, H. J. (1966). *The Sixteen Personality Factor Questionnaire*, 3rd Edition. Instit. Pers. and Abil. Testing, Champaign, Illinois.
- Cattell, R. and Warburton, F. (1967). *Objective Personality Measurement: General Principles and a Compendium of Tests*. Champaign, Illinois; Univ. of Illinois Press.
- Cooley, W. W., and Lohnes, P. R. (1962). *Multivariate Procedures for Behavioral Sciences*. New York: Wiley and Sons.
- Degan, J. W. (1952). *Dimensions of Functional Psychosis*. Richmond: Byrd Press.
- Eades, D. C. (1964). General biological and geographic variation of *centrophilus guttulosus* Walker (orthoptera: gryllacrididae: raphidophorinae). *Trans. Amer. Entom. Soc.*, 90, 73-110.
- Flament, C. (1963). *Applications of Graph Theory to Group Structure*. New York: Prentice Hall.
- Holzinger, K. J., and Harman, H. (1941). *Factor Analysis*. Chicago: Univ. of Chicago Press.
- Horn, J. L. (1961). Significance tests for use with  $r_p$  and related profile statistics. *Educ. psychol. Measmt.*, 2, 363-370.
- Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart and Winston.
- Howard, K. and Diesenhaus, H. (1965). Direction of measurement and profile similarity. *Instit. Juv. Res., Chicago Res. Repts.*, 2, 7, 1-24.
- Jane's Fighting Ships (1964-65). London: Jane's Fighting Ships Pub. Co.
- Jung, C. G. (1923). *Psychological Types*. London: Routledge and Kegan Paul.

- Kaiser, H. F., and Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, 30, 1-14.
- Lazarsfeld, P. F. (1960). Latent structure analysis and test theory. In H. Gulliksen and S. Messick (eds.), *Psychological Scaling: Theory and Applications*. New York: Wiley.
- Lorr, M. (1962). Measurement of the major psychotic syndromes. *Ann. NY Acad. Sci.*, 93, 851-856.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci., Calcutta*, 12, 49-55.
- McQuitty, L. L. (1963). Rank order typal analysis. *Educ. psychol. Measmt.*, 23, 55-61.
- McGee, V. E. (1965). The multidimensional analysis of elastic distances. Rept. on ONR Grant NONR-3897-05. Dartmouth College, Hanover.
- Nunnally, J. (1962). Analysis of profile data. *Psychol. Bull.*, 59, 311-319.
- Overall, J. E., and Hollister, L. E. (1964). Computer procedures for psychiatric classification. *J. Amer. med. Assoc.*, 187, 583-588.
- Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: Wiley and Sons.
- Rogers, C. R. (1951). *Client-centered therapy*. New York: Houghton Mifflin.
- Shepard, R. N. (1962). The analyses of proximities: multidimensional scaling with an unknown distance function. I and II. *Psychometrika*, 27, 125-139, 219-246.
- Sokal, R. R., and Sneath, H. A. (1963). *Principles of Numerical Taxonomy*. New York: Freeman and Co.
- Stephenson, W. (1936). The inverted factor technique. *Brit. J. Psychol.*, 26, 344-361.
- Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: Univ. of Chicago Press.
- Tucker, L. R. (1963). Implications of factor analysis of 3-way matrices analysis for the measurement of change. In C. W. Harris (ed.), *Problems in Measuring Change*. Madison: Univ. of Wisconsin Press.
- Wittenborn, J. R. (1951). Symptom patterns in a group of mental hospital patients. *J. consult. Psychol.*, 15, 290-302.

**Table 1. Values from Distribution of Random  $r_p$ 's Obtained by Monte Carlo Methods**

(Using normal distribution on each element of profile.)

Algebraic Mean				
N/k	2	6	10	
25	.188	.135	.052	
50	.174	.047	.011	
75	.114	.046	.017	
100	.100	.031	.011	

Mean of Positive Values Only				
N/k	2	6	10	
25	.501	.307	.223	
50	.479	.263	.197	
75	.450	.263	.202	
100	.449	.257	.199	

**Table 2. Incidence Matrix for 15 People**

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	1	1	1	1	1	1	1	1	1						
b	1	1	1	1	1	1	1	1	1						
c	1	1	1	1					1	1					
d	1	1	1	1					1	1					
e	1	1			1	1	1								
f	1	1			1	1	1								
g	1	1			1	1	1								
h	1	1	1	1				1							
i	1	1	1	1	1				1						
j										1	1			1	
k										1	1	1		1	1
l											1	1	1	1	1
m												1	1		1
n										1	1	1		1	1
o											1	1	1	1	1

Table 3. Process Sequences in the Boolean Algorithm for Phenomenal Cluster Search

(1)

	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	1	0	1	0	0	0
2	0	1	0	0	0	1	0	0	0	0
3	0	0	1	0	0	0	1	0	0	0
4	0	0	0	1	0	0	0	0	0	0
5	1	0	0	0	1	1	1	0	0	0
6	0	1	0	0	1	1	1	0	0	0
7	1	0	1	0	1	1	1	0	1	0
8	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	1	0	1	1
10	0	0	0	0	0	0	0	0	1	1

 $Q_0$ 

(a)

I	II	III	IV	V		1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	0	I	1	0	0	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0	0
2	0	1	0	0	0	II	0	1	0	0	0	1	0	0	0	2	0	1	0	0	0	1	0	0	0	0
3	0	0	1	0	0	III	0	0	1	0	0	1	0	0	0	3	0	0	1	0	0	0	1	0	0	0
4	0	0	0	0	0	IV	0	0	0	0	0	1	0	1	0	4	0	0	0	x	0	0	0	0	0	0
5	1	0	0	0	0	V	0	0	0	0	0	0	0	1	1	5	1	0	0	0	1	x	1	0	0	0
6	0	1	0	0	0											6	0	1	0	0	x	1	x	0	0	0
7	1	0	1	1	0											7	1	0	1	0	1	x	1	0	1	0
8	0	0	0	0	0											8	0	0	0	0	0	0	0	x	0	0
9	0	0	0	1	1											9	0	0	0	0	0	0	1	0	1	1
10	0	0	0	0	1											10	0	0	0	0	0	0	0	0	1	1

$G_1$        $G'_1$       =       $Q_1$

(b)

Table 3 Continued

5.31

(3)

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	1	1	1	0	0	0
6	0	0	0	0	1	1	1	0	0	0
7	0	0	0	0	1	1	1	0	0	0
8	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

 $Q_2$ 

(c)

	I	II	III	IV	V	VI
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	0	0	0
(4) 5	1	0	0	0	0	1
6	0	1	0	0	0	1
7	1	0	1	1	0	1
8	0	0	0	0	0	0
9	0	0	0	1	1	0
10	0	0	0	0	1	0

 $G_2$ 

	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	1	0	1	0	0	0
2	0	1	0	0	0	1	0	0	0	0
3	0	0	1	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	1	1	1	0	0	0
6	0	1	0	0	1	1	1	0	0	0
7	1	0	1	0	1	1	1	0	1	0
8	0	0	0	0	0	0	0	x	0	0
9	0	0	0	0	0	0	1	0	1	1
10	0	0	0	0	0	0	0	0	1	1

 $Q_3$ 

(d)

Table 4. Phenomenal Clusters Discovered

Row Size	Identifying Number	Instances	Size, as Percentage of Sample	Frequency Distribution
6	(2)	a b c d e i	60	16 2/3
5	(1)	a b c f g (3) a b c d h	50	33 1/3
4	(4)	k l n o	40	16 2/3
3	(5)	j k n (6) l m o	30	33 1/3
2	None		20	0

Table 5. Account of Nuclear Clusters

Size	Phenomenal Clusters Involved			Size, as Percentage of Sample		Frequency Distribution	
	2	3	4	2	3	2	3
4	a b c d			40		16 2/3	
3	e a b			30		16 2/3	
2	c d; k n; o,	a b		20	20	50	100
1	e			10		16 2/3	

Table 6. Finding Segregates by the Continuous Connectedness Algorithm

(a) Phenomenal Cluster Contiguity Matrix,  $Q_c$ 

Phenomenal Cluster Identifying Numbers	1 (6)	2 (5)	3 (5)	4 (4)	5 (3)	6 (3)
1 (6)	6	3	4	0	0	0
2 (5)	3	5	3	0	0	0
3 (5)	4	3	5	0	0	0
4 (4)	0	0	0	4	2	2
5 (3)	0	0	0	2	3	0
6 (3)	0	0	0	2	0	3

Entries state the count of overlap of persons.

## (b) Incidence Matrix among Phenomenal Clusters

Phenomenal Cluster Identifying Numbers	PQ					
	1	2	3	4	5	6
1	1	1	1	0	0	0
2	1	1	1	0	0	0
3	1	1	1	0	0	0
4	0	0	0	1	1	1
5	0	0	0	1	1	0
6	0	0	0	1	0	1

Converted to Incidence Matrix for 2 overlap and above.

## (c) Segregates Discovered by Segregate Search Algorithm Applied to (b).

	$S_1$	$S_2$	
1	1	0	
2	1	0	
3	1	0	$S_1 = a b c d e f g h i$
4	0	1	
5	0	1	$S_2 = j k l m n o$
6	0	1	

Table 7. Boolean Algorithm for Continuous Connectedness Search

(a)			(b)			(c)		
0	1	0	0	0	0	0	0	0
1	0	0	1	1	1	1	1	1
1	0	0	1	0	0	1	0	1
0	1	0	0	1	0	0	1	1
0	1	0	0	1	0	0	1	1

Table 8. Nuclear Types Found Among Nations by Culture Pattern

$t_p$  Evaluations

## Eastern European

Czechoslovakia  
Estonia  
Lithuania  
Austria

## Mohammedan

Afghanistan  
Iraq  
Turkey  
Arabia  
Egypt

## Scandinavian

Denmark  
Sweden  
Norway  
Switzerland

## Commonwealth

New Zealand  
Australia  
Netherlands  
Belgium  
Canada

## Oriental

India  
China  
Tibet



Table 9. p-Cluster Search Stage of Taxonome Illustrated on Jane's Fighting Ships

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
	Carriers Destroy. Submarines										Frigates																		
1	1	0	1	1	1																								
2	0	1	1	1	0																								
3	1	1	1	1	0																								
4	1	1	1	1	1																								
5	1	0	0	1	1																								
6						1	1	1	1																				
7						1	1	1	1																				
8						1	1	1	1																				
9						1	1	1	1																				
10										1	1	1	1	1	1	1	1	1	1	1									
11										1	1	1	1	1	1	1	1	1	1	1									
12										1	1	1	1	1	1	1	1	1	1	1									
13										1	1	1	1	1	1	1	1	1	1	1									
14										1	1	1	1	1	1	1	1	1	1	1									
15										1	1	1	1	1	1	1	1	1	1	1									
16										1	1	1	1	1	1	1	1	1	1	1									
17										1	1	1	1	1	1	1	1	1	1	1									
18										1	1	1	1	1	1	1	1	1	1	1									
19										1	1	1	1	1	1	1	1	1	1	1									
20																				1	1	1	1	1	1	1	1	1	
21																				1	1	1	1	1	1	1	1	1	
22																				1	1	1	1	1	1	1	1	1	
23																				1	1	1	1	1	1	1	1	1	
24																				1	1	1	1	1	1	1	1	1	
25																				1	1	1	1	1	1	1	1	1	
26																				1	1	1	1	1	1	1	1	1	
27																				1	1	1	1	1	1	1	1	1	
28																				1	1	1	1	1	1	1	1	1	
29																				1	1	1	1	1	1	1	1	1	

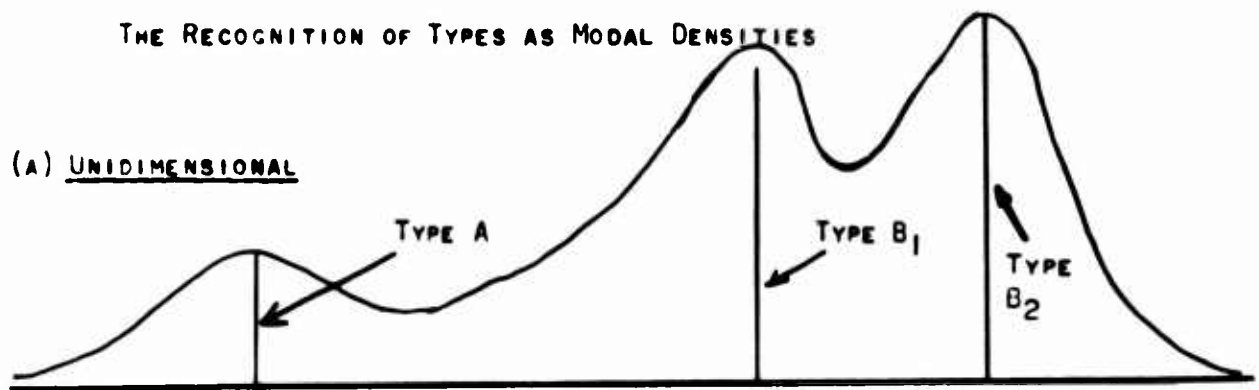
## p-Clusters after One Cycle

	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	1	1	0	0
2	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	1	0	1	0
4	0	0	0	0	0	1	1	1	0
5	0	0	0	0	1	0	1	0	0
6	0	0	0	1	0	0	0	0	0
7	0	0	0	1	1	0	0	0	0
8	0	0	0	1	1	0	0	0	0
9	0	0	0	1	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	1
13	1	0	0	0	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0
17	1	0	1	0	0	0	0	0	1
18	1	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0
20	0	1	1	0	0	0	0	0	0
21	0	1	0	0	0	0	0	0	0
22	0	1	1	0	0	0	0	0	0
23	0	1	0	0	0	0	0	0	0
24	0	1	0	0	0	0	0	0	0
25	0	1	0	0	0	0	0	0	0
26	0	1	0	0	0	0	0	0	0
27	0	1	1	0	0	0	0	0	0
28	0	1	0	0	0	0	0	0	0
29	0	1	1	0	0	0	0	0	1
	1	0	1	1	1	0	0	0	0

DIAGRAM 1

THE RECOGNITION OF TYPES AS MODAL DENSITIES

(A) UNIDIMENSIONAL



(B) BI- OR MULTI-DIMENSIONAL

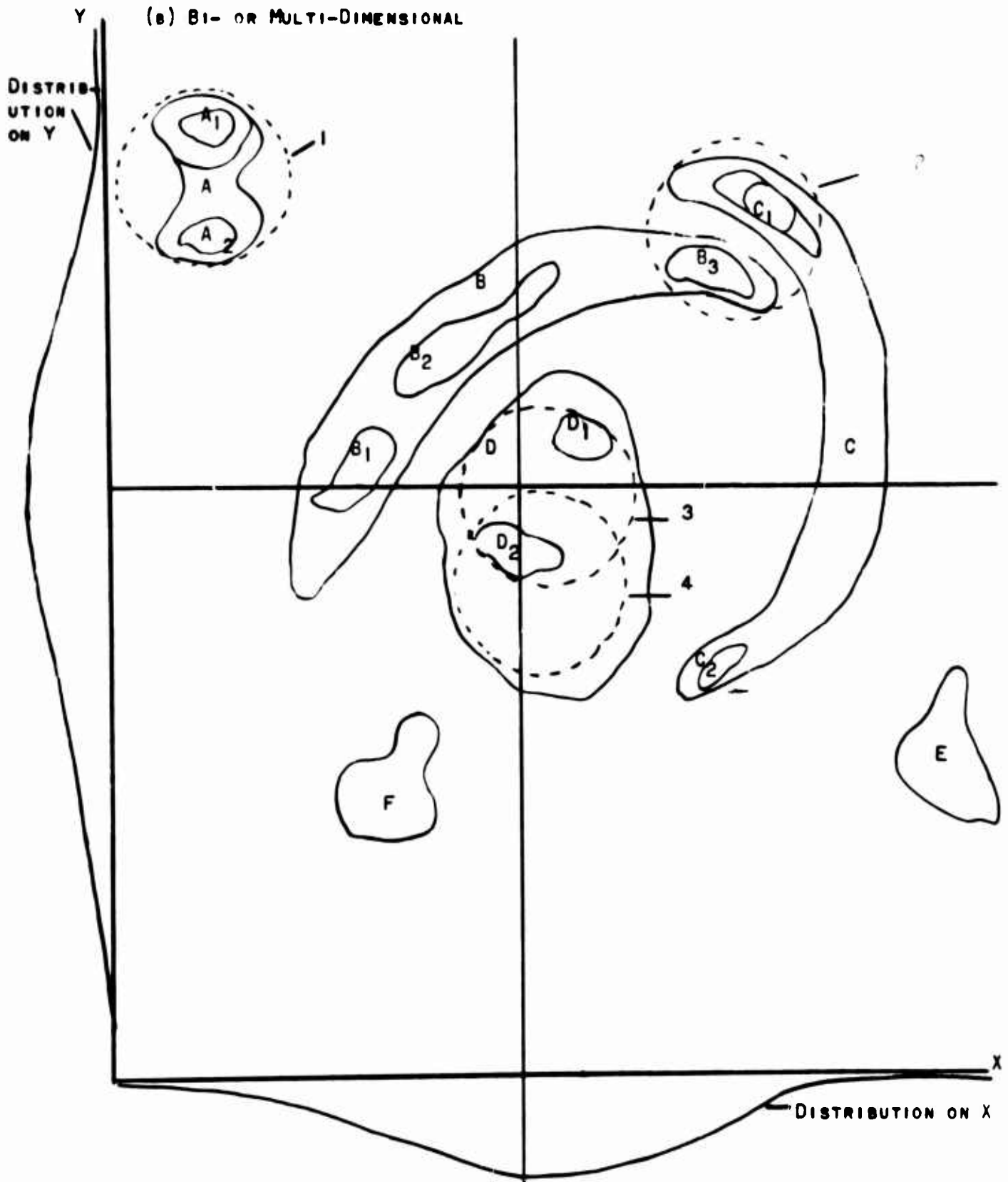
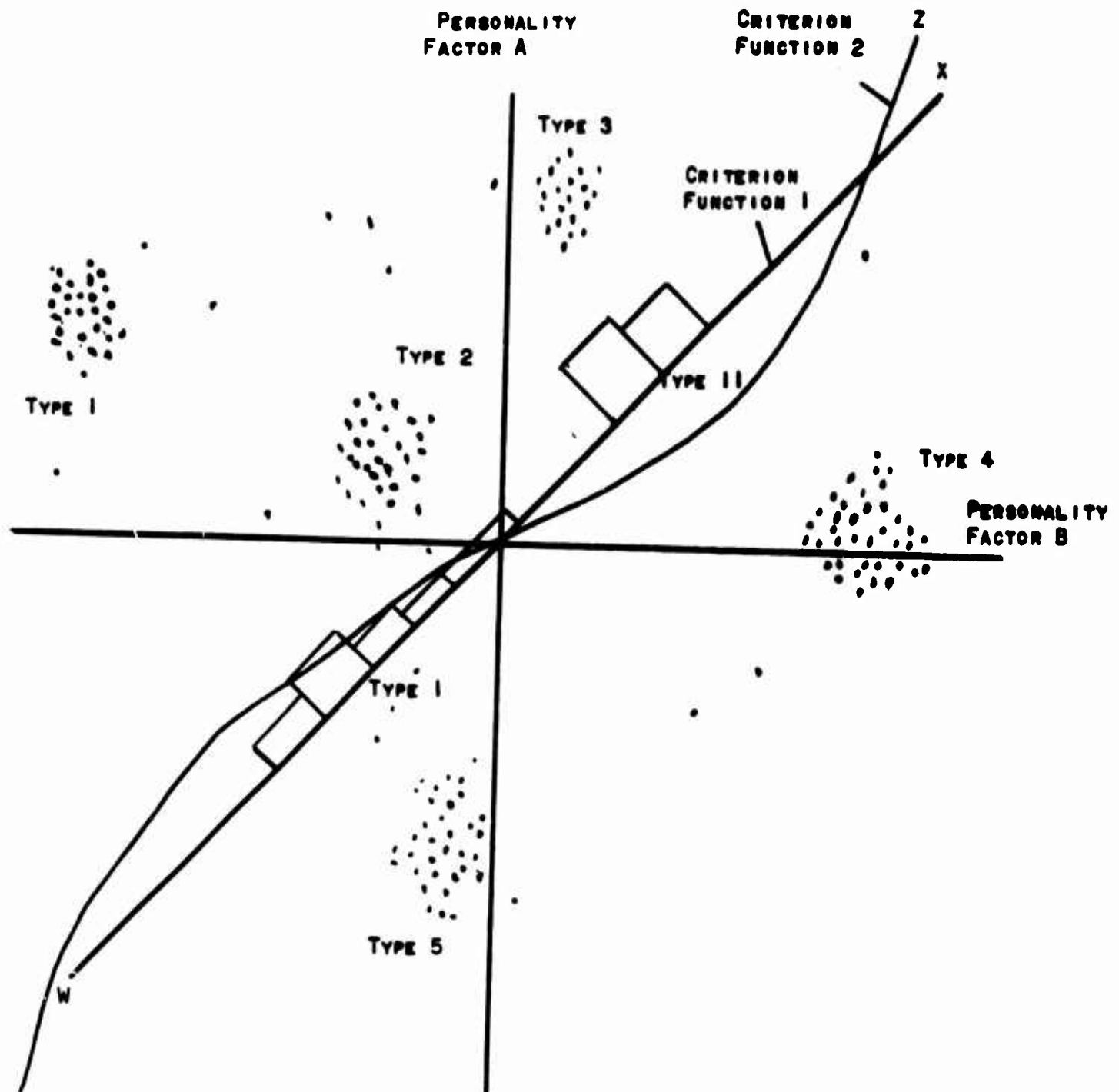


DIAGRAM 2

REGROUPING AND REDUCTION OF MODAL  
TYPES WHEN SPECIFIC *EFFECT* (CRITERION) FUNCTIONS  
ARE USED



Y HISTOGRAM DRAWN OF CRITERION FUNCTION 1 SHOWS TWO MODES (TYPES)  
REPLACING THE FIVE MODES (TYPES) ON THE PROFILE ELEMENTS THEMSELVES.  
THE PROJECTIONS ON CRITERION FUNCTION 2 ARE TOO COMPLEX TO BE READILY DRAWN.

Comparative Cluster Analysis of Variables and Individuals:  
Holzinger Abilities and the MMPI<sup>1</sup>

Robert C. Tryon  
University of California, Berkeley

The first objective in comparative cluster analysis is to describe the similarity of the dimensions discovered in different groups. This problem is known as the comparative dimensional analysis of variables, or "factor-matching". In the domain of the intellectual abilities, for example, one may discover in a middle-class suburban group of children that the 24 Holzinger tests of diverse specific abilities (Holzinger and Swineford, 1939) can be accounted for by four "basic" general abilities, or factors, Verbal, Space (Form), Speed, and Memory, symbolized as V, S, P and M. Are these dimensions identical with those found in a lower-class school of children of factory workers? An MMPI example: Are the seven general dimensions of Introversion, Body, Suspicion, Tension, Depression, Resentment and Autism found in a group of psychiatric patients the same ones discovered in a group of normals?

This problem has a direct, simple solution when approached by the logic and procedures of cluster analysis based upon domain sampling principles and incorporated procedurally in the BC TRY Computer System of cluster and factor analysis (Tryon and Bailey, 1965). Since dimensional analysis requires as basic data the intercorrelations between the variables in the groups, you might reasonably ask this question: How can one compute the correlations between variables of different groups of subjects? The answer is

that in comparative dimensional analysis all that is needed are the factor coefficients of the dimensions within each group (these being referred to in factor analysis as the "rotated oblique factor coefficients"). These factorial findings within the different groups are directly compared across the groups by the comparative cluster analysis programs called COMP1 and COMP2 of the BC TRY System.

The second general objective is that of comparing the typologies of two or more groups of individuals. When, for example, we score the Factory and the Suburban subjects on the four general abilities, V, S, F and M, we can objectively sort the children in each group into different types based upon the patterns of their scores on V, F, S, and M. These person-clusters (or profile types) in the two groups can differ in two ways. First, even though the same kinds of profile types may appear in the two groups, those that occur with high frequency in one group may be rare in the other group. We may refer to this type of typological similarity across groups as the similarity of their "frequency-patterns" on a common typology. Second, the kinds of types in the two groups may be different; those that compose one group may not match the types of the other group. In the BC TRY System, the programs expressly designed to perform the comparative typology of groups are the components OTYPE, OSTAT and EUCC.

The plan of this paper is as follows: The comparison of the dimensions of different groups (COMP) and of their typologies (OCOMP) will first be made for the case of the Holzinger study of the abilities of two groups, the Factory and the Suburban children. Under exactly the same format of analysis you will then find the COMP and OCOMP analysis of the Patient and the Normal groups in a study of MMPI item-clusters. Our interest in these two studies is as much substantive as procedural, because they refer to two important problems in cognitive and personality psychology.

## The Study of Abilities: The Holzinger Problem

### Comparative dimensional analysis ("matching factors" or COMP analysis)

The 24 variables.--In the Holzinger problem, 301 grade school children were given 24 separate tests of specific abilities. These tests are listed in Table<sup>6</sup>1, where you will note that they are grouped under the five domains of Spatial, Verbal, Speed, Memory and Mathematical. Most of these tests may be recognized as forms that are included today in test batteries of "Intelligence", such as, for example the WISC and WAIS batteries from which the Verbal, Performance and Full Scale IQs are determined (Anastasi, 1961, Chapter 12).

The groups.--The total group of children, here called the Inclusive group, were children from two Chicago grade schools. The authors (Holzinger and Swineford, 1939, p.6) describe them as follows: "The children in the Pasteur School came largely from the homes of workers in near-by factories. Many of the parents were foreign-born....using their native language at home....Both parents were American-born in 29 per cent of the cases, while in 48 per cent, both were foreign-born." The second school was the Grant-White school in the suburb of Forest Park, Ill. In this group "...both parents were American-born in 72 per cent of the cases while both were foreign-born in only 15 per cent. Almost 100 per cent of the children were born in the suburb in which the school was located."

The Inclusive Group can therefore be thought of as being composed of two ecological groups. The 156 from the Pasteur School will be here called the Factory Children, the 145 from the Grant-White School, the Suburban Children. The data from this last Suburban group have been made famous as a basic data-set in factor analysis history, being known as "The Holzinger-Harman Problem" (Harman, 1960). The Inclusive Group has other subgroup structures, notable sex groups and grade groups. Furthermore, the Suburban Children were organized into two types of classrooms, "homogeneous groups" and random classes.

Dimensional analysis of the 24 variables in the Inclusive Group.--A direct comparison of the dimensions of the 24 variables in the Factory and Suburban Children and of their separate typological structures can only be made when the definers of their dimensions are the same. The first objective, therefore, is to decide on the number of dimensions on which the subgroup comparisons are to be made, and on a common set of definers of each dimension. A full-cycle key cluster analysis of the 24 variables (Tryon & Bailey, 1965, Table 1, Section B) was performed on the Inclusive group, from which it was discovered that after four dimensions were extracted from the intercorrelations among the 24 tests, their residuals were trivial. Many different varieties of factor analysis have been performed on the correlations of the Suburban Children, all of which also find four salient dimensions (Harman, 1960).

Table 1  
about  
here

The defining variables of each of the four dimensions are shown by superscripts attached to the names of the variables in Table 1. Thus, under the Spatial category all four of the spatial tests are marked with super "f", indicating that each is a definer of one dimension, the F dimension, measuring form (or space) perception. Analogously, four "v" tests define the V or Verbal, four "s" the S or Speed, and five "m" the M or Memory dimensions. No fifth dimension was required for the mathematics tests.

Dimensions V, S, F, and M are thus designated as the "basic" general dimensions of the 24 abilities, on which the comparative dimensional and typological analyses of subgroups are performed. Details on the dimensional analysis of the Inclusive group are not given here for two reasons: They have been recently published elsewhere (Tryon, 1966b), and they are so similar to those of the Factory Children which are given below (see Fig. 1, bottom) that no useful purpose is served by presenting them.

Dimensional analysis of the 24 variables in the Factory Children.--To discover the cluster structure of the tests in the Factory Children, a full-cycle key-cluster solution of this group's intercorrelations among the 24 tests was "preset" on the definers of the four basic dimensions found in the Inclusive Group.



The results are shown pictorially in Fig. 6.1, the bottom spherical plot, which is a direct tracing of the printout of the diagram in program SPAN (SPherical ANalysis) of the BC TRY System. The surface separation of any two tests on this sphere is a function of the correlation between them (technically, of their "inter-domain", or "common-factor correlation"). Two tests that correlate unity have superimposed points, two that correlate zero are  $90^\circ$  apart, represented in Fig. 6.1 by the distances between the three boxes that form the spherical triangle; the boxes represent the subset of three independent dimensions derived by factoring on residuals.

Note in Fig. 6.1 that the five Verbal tests cluster tightly together at lower left in the configuration, the four Speed tests more loosely at lower right, the four Form tests at the top. The six Memory tests are marked by "X", denoting that they all project into a fourth dimension which cannot be shown since it projects at right angles to the three depicted in Fig. 6.1. Note, however, that the five mathematical tests are depicted in these three dimensions, and that they are all "dependent" on V, S, F and M in the sense of being predictable from the four, a point proved in a recent paper (Tryon, 1967b).

For readers in whom the thought may lurk that this clear cluster structure is due to "presetting" on the definers of the Inclusive group, it is regrettable that space does not permit showing the configuration recovered by a purely blind empirical key-cluster factoring of the Factory correlation matrix. To do so would, however, be redundant because the empirically-derived configuration differs only trivially from that shown in this preset solution. The same configuration also results from an orthodox principal-axes solution plus varimax or quartimax rotation, also available in the BC TRY System. Indeed, the same configuration is necessarily the same for all varieties of factoring on a given set of dimensions that result in trivial residuals.

Dimensional analysis of the 24 variables in the Suburban Children. --Applying the same dimensional procedure to the

correlation matrix of the Suburban Children gives, as a result, the configuration shown in the top SPAN diagram of Fig. 1. At lower left in the configuration is the same Verbal cluster as in the Factory group; at lower right Speed, at the top Space, and the Memory cluster also projects into a fourth dimension; the mathematical abilities once again deploy centrally as dependent variables predictable from the V, S, F, and M dimensions. Clearly the cluster structure of the Suburban Children closely resembles that of the Factory. One obvious difference is that, though the cluster groups are about the same, they are, as groups, more separated from each other in the Factory than in the Suburban Children, that is, less correlated with (oblique to) each other.

Comparison of the dimensions within each group separately (COMP1). --A metric description of the within-group structures is provided by a program that computes the correlations between the ability clusters defined as oblique dimensions, computed by the CSA (Cluster Structure Analysis) program of the BC TRY System. The values of these correlations are given in Table 2, section A, where you see the correlation matrix of the V, S, F, and M dimensions. These correlations are known in factor analysis as the "correlations between rotated oblique factors", or their "common factor correlations". In cluster analysis they are called "inter-domain" correlations, where each cluster is conceptualized as a domain score,  $C_i$ , on many variables collinear with the observed definers of the cluster (Tryon, 1959, equation 24). Thus, the domain score,  $C_v$ , on the Verbal cluster is a hypothetical score on many variables collinear with the observed set,  $V_5, V_6, V_7, V_8$ , and  $V_9$ , shown in the SPAN diagram. (The term "collinear" means projecting to the same degree on the same vector from the origin of the sphere.)

The inter-domain correlations, listed in Table 2 under the columns headed  $r_{CC}$  are computed from the raw correlation matrix using the well-known formula for the "correlation of sums". As you look through the  $r_{CC}$  values you find precise metric expressions of the degree of similarity of the four basic ability-dimensions, V, S, F and M, in the Suburban and Factory Children.

Table 2  
about  
here

For example, the inter-domain  $r_{CC}$  between the Verbal and Speed dimensions is seen to be .63 and .42, respectively; that is, the two dimensions have almost exactly the same degree of similarity in the two groups. But between the other dimensions you will find that the  $r_s$  are generally higher for the Suburban than for the Factory children, a fact already seen visually in the SPAN diagrams of Fig. 1. The  $r_{CC}$  values are thus a metric statement of similarity that is displayed visually on the spheres.

In the lower sections of Table 2, you will find other metric properties of the four basic ability dimensions. The "generality" of each, given in section C, is the degree to which each dimension accounts for all the raw inter- $r_s$  among the 24 abilities. In both groups the Verbal dimension is the most general, but in the Factory group the other three dimensions are more specific than in the Suburban. Of special interest to the typological analysis is the reliability coefficient of the raw scores on the four dimensions. In section D, the reliability of V is .9, but of the other three, only of the order .7 or .8. (The formula for reliability is known as alpha, though a better term is the Variance Form (Tryon, 1957).)

Direct comparative analysis of the dimensions across groups (COMP2).—To this point we have assessed the similarity of the V, S, F and M dimensions of the Factory and Suburban Children by the subjective process of cross-referencing their separate configurations in Fig. 1, and by comparing their within-group  $r_{CC}$  values in Table 2, procedures that are indirect and inferential. Can we directly compare their dimensions?

Fig. 2  
about  
here

Fig. 2 displays the direct comparison achieved by the program COMP2 of the BC TRY System. In this SPAN diagram, traced from the printout, you will note that the Verbal dimension of the Suburban Children, labelled  $V_G$  (for the GW school) and that of the Factory Children, labelled  $V_P$  (for Pasteur) are tightly clustered at lower left, meaning that they are quite similar. At lower right are the two points representing the Speed dimensions of the two schools; at the top you see their two Space dimensions, and extending into the fourth dimension are their two Memory

dimensions. This cluster structure therefore directly compares in one diagram the similarity of the two-dimensional structures that we only indirectly observed above by cross-referencing.

The direct index of the similarity of any two dimensions across different groups is the "index of similarity" of the two dimensions (or "factors"), called the cos  $\theta$  between them. For two dimensions within a group cos  $\theta$  is equivalent to the inter-domain correlation,  $r_{CC}$ , but it is estimated not from the raw correlation matrix, as is  $r_{CC}$ , but from the oblique factor coefficients of the two dimensions. The proof that cos  $\theta$  between two dimensions within a group is  $r_{CC}$  is given in Table<sup>6</sup>2, section A; there you will find in the columns labelled "Cos  $\theta$ " this index of similarity (computed by COMP2) set beside the  $r_{CC}$  value (computed by program CSA). You will find that the two indices are virtually identical in every case.

But since the similarity index, cos  $\theta$ , is computed only from factor coefficients, it can be, of course, calculated for dimensions across different groups. These similarity values are given in Table<sup>6</sup>2, section B. They tell the same story metrically that is shown pictorially in the spherical configuration of Fig.<sup>6</sup>2. On the upper left to lower right diagonal you see the index of similarity of V in the Factory and in the Suburban Children, then of S, F, and M. For example, that between the Verbal dimensions in the two groups is .96, between the two Speeds it is .89, between Forms .92, between Memories .83.

For the technically-minded reader I include in Appendix A the logic and formulation of cos  $\theta$  as an index of dimensional similarity. Briefly, the reasoning by which we designate two dimensions as identical is based on the universal logic by which we conceive any two entities as being the same, namely, that they show the same pattern of observations in relation to a common set of other "referent entities". For example, the Verbal dimensions in the two groups are virtually identical because their patterns of factor coefficients (the observations) on the constant set of 24 referent abilities are virtually identical. The index of pattern similarity of any two entities on a common set of referent entities is P, called the index of proportionality,

or collinearity, described in detail in Appendix A for the case of pattern similarity of the factor coefficients of any two dimensions. The value of the index of similarity,  $\cos \theta$ , of any two dimensions in different groups is a simple quadratic function of  $P$ , as shown in Appendix A.

To sum up, we find in Fig. 6.2, and from the metric values in Table 2 that the four basic dimensions V, S, F, and M in the two groups are highly similar. But in the Factory Children they are somewhat more independent of each other than in the Suburban. Why? An environmental explanation is that the parents of the Suburban Children stress scholastic achievement, implementing their ambition by pushing their "promising" children in all abilities, letting their less promising children fend for themselves. Consistent with this theory, we find that it is precisely in the Suburban Children that the scholastic institution of "homogeneous" classification is employed, namely, the sorting of sheep and goats into different classrooms. In the Factory group, children generally are left to fend for themselves.

But there is an alternative genetic explanation: There probably is more stringent assortative mating on abilities among Suburban parents. This sort of sexual selection would generate a higher correlation among all abilities in the Suburban group than in the Factory, where assortative mating would be more random. A systematic treatment of such environmental vs. genetic "correlation-producing" agencies in the case of abilities is presented elsewhere (Tryon, 1935, 1939).

#### Comparative typological analysis in the Holzinger Problem (OCOMP analysis).

When we allocate children having the same patterns of scores on the basic abilities, V, S, F and M, to O-types, do we find the same typological structure of these O-types in the Factory and Suburban groups?

Similarity of frequency-patterns of the two groups on the common typology of the Inclusive group.--The first of two ways of determining the typological similarity of two groups is to discover

Table<sup>6.3</sup>  
about  
here

the degree to which they show the same frequency of cases falling in the common typology of the Inclusive Group. You will find this common set of O-types in Table<sup>6.3</sup> under the general heading at left, "Inclusive typology". (How these types are determined will be described later.) Look at the first type, labelled H1, consisting of 14 children whose pattern of cluster scores on basic abilities, V, S, P, and M is 48, 36, 44, 37, respectively. These are mean standard Z-scores on a scale whose mean for the 301 children in the Inclusive Group is 50, and sigma 10. Underlined scores of 40 (-1 sigma) or below are termed "Low" in the column headed "Descriptive name", those 60 (+1 sigma) or above are called "High". For this H1 type you will therefore find it described in the table as "Low Speed and Memory".

As you look down the column of types from H1 through H15 to the class called Unique you see in the adjacent frequency column that some types have a high frequency, like H9, the Average type, with 38 children in it, others with low frequency, like H2, the Low Verbal and Memory type, with only eight cases in it. Our logic of typological similarity of the Factory and Suburban Children is simply this: If both groups show the same frequency pattern on these common 16 Inclusive classes, then they have the same typological structure, but to the degree that their frequencies in these 16 classes differ from each other their typologies are obviously different.

Before examining the findings, I will briefly review how the typology of the Inclusive group is determined, a matter published in some detail in a recent paper (Tryon, 1967a). The cluster scores of each of the 301 children are first computed by program FACS (Factor And Cluster Scores) by the BC TRY System. For example, their scores on V Verbal are the mean of their standard scores on the four defining variables of V (listed in Table<sup>6.1</sup>), restandardized on a scale of mean 50, sigma 10. The program OTYPE inputs the 301 cluster scores and completely objectively allocates them to the 16 classes given in Table<sup>6.3</sup>. The principles of classification, called the Condensation Method, are quite simple: All 301 scores are located as points in the cluster score space of the four dimensions defined by the V, S,



$P_i$  and  $M$  scales, and, working from the Euclidean distances between them, the program allocates those with small distances between them, that is, those in a tight object-cluster in this space, into the same  $O$ -type. Another program, OSTAT (Object STATistics), then computes the mean  $Z$ -scores of the individuals in each cluster, given in Table 6.3, and also calculates an index of homogeneity,  $H$ , or tightness of the cluster (For details, see Tryon, 1967a).

Turn now to the similarity of the frequency-patterns of the Factory and Suburban groups. From the OSTAT printout mentioned just above, it is a simple matter to count how many children in each group fall into the 16 classes, from which the percentage falling into each class is computed. These percentages are printed in Table 6.3 under the general heading "Factory vs. Suburban". The listed values in the two columns labelled " $p_F$ " and " $p_S$ " are the critical frequency-patterns of the two groups, on the basis of which their typological similarity is determined. The overall index of similarity, given just below the table, is the same general index of proportionality,  $P$ , discussed earlier, the formula for which is printed below the table. If you inspect this formula, you will discover that if two groups have exactly the same frequency-patterns, i.e.,  $p_F = p_S$ , then the index  $P$  is unity (1.00). But if their patterns are utterly different, that is, if the occurrence of each type in one group is matched with the absence ( $p = 0$ ) in the other group, then the index,  $P$ , is zero. I have worked out the value of  $P$  for the two ecological groups of children below the table, where you will find it to be  $P = .75$ , denoting a considerable amount of typological similarity of the two groups.

Of greater interest, however, are the specific type differences between the two groups. These values are listed under "Diff" in Table 6.3. Because the sampling error of such differences can be large, it is desirable to indicate which of these differences is unlikely to occur by chance at the strong confidence level of  $p < .001$ . Fortunately, we are working with small values of per cents which keep the error down. Expressing the per cents as proportions,  $p_i$ , we note that the mean proportion in the 16 classes is  $\Sigma p_i / 16 = 1.00 / 16 = .06$ . Since most of the

proportions of the types in both groups are not too greatly different from .06, we will compute the standard error of a difference between two true proportions of .06, using the well-known formula for this error printed at the bottom of the table, and worked out for the Ns of the two groups. It comes to .3. We may therefore set a per cent of 3 as the lower bound at and above which any difference is almost surely non-chance.

You will find all differences above 3 indicated by an (S) for Suburban or an (F) for Factory, depending on which group has the highest per cent. For example, note that the largest difference between per cents is 12 in type H8, Low Verbal. For this difference the greatest per cent frequency is 13 in the Factory group. Next comes H10, Hi Verbal, most characteristic of the Suburban group. These two Verbal types therefore represent the greatest typological difference between the two groups. If you look through the other significant differences you will discover that the Suburban group falls more heavily into Low Memory (H3) and Low Speed (H7) whereas the Factory Children occur more frequently in the Hi Memory (H14) and Hi Speed (H11) types. Verbal, Memory and Speed therefore most markedly differentiate the typological differences between Factory and Suburban children.

Since sex differences in abilities are of universal interest, I have also presented the data for determining the typological similarity of the Boy vs. Girl subgroups, in the far right columns of Table 63. From their columns of per cents in the 16 classes, the index of similarity for the sex groups, worked out below the table, is seen to be  $P = .85$ , somewhat higher than for the Factory and Suburban groups. If you examine in detail the significant differences, you will find that boys more frequently fall into Low Speed, Low Memory, and Low Verbal types, the girls being, conversely, in the Hi types in the abilities. On the other hand, girls fall more frequently into Low Form (Space) types, boys into Hi Form. This finding on the Verbal favoring girls, the Form (or Space) favoring boys has been confirmed in many studies, but Low Speed and Low Memory in boys is a less well-known finding.



Similarity of empirically-derived typologies of the groups.

The above analysis informs us of differences between Factory and Suburban Children only on the single common typology of the Inclusive group. But for fuller information, we need to discover empirically the typology of each group independently of the other, and to compare directly their two typologies. The procedures for doing so are available in programs of the BC TRY System. On the 156 Factory Children separately we objectively determine their typology by the OTYPE and OSTAT programs described above. You will find its 15 classes in Table<sup>6</sup>4, where under "Factory Children" they are listed as types F1 down through F14 to Unique. Their Z-score profile values and descriptive names are also given. You will also find their homogeneity, or  $\bar{H}$  coefficients that describe how "tight" each O-type is in its Z-scores on the four dimensions. This coefficient has been described in detail elsewhere (Tryon, 1955, 1967a) and with special emphasis in a recent paper on the prediction of "outside" attributes of O-types (Tryon, 1967b). Suffice here to say that an  $\bar{H}$  value of 1.00 means that all individuals in an O-type have exactly the same scores on each of the four dimensions, whereas an  $\bar{H}$  of .00 means that the scores are as variable in all four dimensions as is the full supply of all 301 children.

In similar fashion the separately worked-out typology of the Suburban Children is given at the right in Table<sup>6</sup>4, where you will find the 13 classes of these children listed from S1 through S12 to Unique.

You can get a general impression of the typological similarity of the two groups by comparing the descriptive names of the two and by noting from these names which types are present in both groups and which ones are present in one but absent in the other.

We need a more precise comparison of the different typologies. To achieve such precision we project all the 26 types of both groups (14 F types plus 12 S types) into the same analysis, from which we get exact values of the similarities and differences between them. The procedures for doing so are called "EUCO-analysis" in the BC TRY System. The logic of the analysis is quite simple: Each type is considered to be an abstract "individual" plotted as a point in the cluster score space of V, S, F, and M where its locus is

Table<sup>6</sup>4  
about  
here

determined by its four Z-scores listed in Table<sup>6</sup>4. Program EUCO computes the Euclidean distance between each pair of types, and prints these values in a pair-comparison matrix from which one can read off precisely the degree of similarity between any two types.

Space limitations do not permit printing this Euclidean distance matrix here. In its stead, however, I present a pictorial representation of the distances between the types in the form of the SPAN diagram given in Fig.<sup>6</sup>3. To secure this diagram, the EUCO matrix is first transformed to a correlation matrix by correlating columns of EUCO values, then running this r-matrix through a standard key cluster analysis, ending in the SPAN diagram of Fig.<sup>6</sup>3.

The configuration on the SPAN sphere describes the similarities and differences between the Factory and Suburban O-types. The circles represent the 14 Factory O-types, the squares the 12 Suburban. I also include in this analysis the 15 Inclusive H-types from Table 3. The sizes of the circles and squares and the length of the underline of the H-types are proportional to the frequency of each type. Note that the four dimensions, V, S, F, and M are also plotted, these being secured by inputting model abstract "individuals" whose four Z-score values are especially selected to enable one to plot the dimension lines as score axes.

The large super-cluster at left center consists of types all in the "LOW" region, meaning that generally they have Z-scores below the mean on all four dimensions. Note, however, that this super-cluster breaks off into two general subclusters. The upper one consists largely of Suburban types S1, S2, S3, fairly well represented by the Inclusive types H1, H4, H3 and H6, whereas the lower subcluster consists largely of F, or Factory types, which with S4, are well-represented by Inclusive types H2, H5 and H7. From these facts we discern the similarities and differences between the types in this general region of low scoring, noting especially that there are real differences in the typologies of the two groups in this region. I leave to the interested reader a detailed study of the rest of the configuration. The scores of O-types can be approximated by reading off projections on the four score axes, but more accurately by reading the actual values and descriptions given in Tables<sup>6</sup>3 and<sup>6</sup>4. Types represented by broken

Fig. 3  
about  
here

circles and squares lie into the fourth dimension.

Generally, a study of the configuration reveals findings similar to those found from the similarity of the frequency-patterns, namely, that Verbal, Memory and Speed most markedly differentiate the Factory and Suburban groups. For example, note at the top of the diagram that Hi Verbal is represented only by a Suburban type, S7. Low Verbal through the southern hemisphere is heavily dominated by Factory types.

A final, salient question is this one: How well do the 15 Inclusive O-types representatively sample the 26 different types in both ecological groups of children? This question is important because in the practical usage of the typology of abilities, these would be the types usually used for the classifications of individuals. The answer is provided by noting whether one or more of the 15 H-types lie in all regions occupied by the 26 Factory and Suburban types. By inspecting the SPAN diagram and by comparing the F and S types of Table<sup>6.4</sup> with the H types of Table<sup>6.3</sup> you will note that the 15 H-types fairly cover the ground.

### The Study of the MMPI

#### Comparative dimensional analysis (COMP analysis)

The second study selected for comparative dimensional and typological analysis is that of the responses to the items of the MMPI by Normals vs. Patients.

The item-variables.--The variables are 118 items of the MMPI drawn from the full item supply of 566 to which the subjects responded. The 118 were those shown in a previous study to be the most salient set (Tryon, 1966b). The method used in the prior study is called the BIGNV procedures of the BC TRY System, a method that enables one to perform cluster or factor analyses unrestricted by the number of variables or number of subjects. The subjects were the Inclusive Group consisting of the Normal and the Patient groups.

The groups.--The Normals were 90 Armed Service Officers

matched for age and education against 220 Patients. The latter were outpatients of a VA Mental Health Clinic, consisting of 70 diagnosed Schizophrenics all with a history of hospitalization within the previous 6 years, and 150 diagnosed Anxieties none with a history of any hospitalization for psychiatric disorder.

**Dimensional analysis of the 118 item-variables in the Inclusive Group.**--Recall from the Holzinger study that a comparative dimensional analysis of two groups, here the Normals and Patients, can only be performed when the subjects are measured on the same dimensions defined by the same variables, usually those discovered in a dimensional analysis of the Inclusive group. This analysis revealed four "basic" MMPI item-clusters: I Introversion, B Body, S Suspicion, and T Tension. The defining items of these four dimensions are those whose item-numbers are listed in Table 5, section A. I do not present a more detailed description of these items because it would be too voluminous; but a paraphrasing of them is given in the previous study (Tryon, 1966b, Table 2), and the exact contents are given in MMPI booklets, generally available to most readers. You will note in Table 5 that each item-cluster consists of a "Full Form" and a "Short Form". The comparative dimensional analysis presented in this section was performed on the scores of subjects on the Short Forms, and it also includes the Short Form items of the other three "dependent" item-clusters, D Depression, R Resentment, and A Autism, whose item-numbers are also given in Table 5, section B.

Table 5  
about  
here

The dimensional analysis of the Inclusive Group from which the four basic and three dependent dimensions were derived cannot be presented here because it is fully explicated in the prior publication. However, the results of it are so similar to those given below on the Patient Group (See Fig. 4, top diagram), that no point would be served in giving the findings here. In sum, it was found that seven dimensions were required to account for the intercorrelations among the 118 items, but that the first three basic dimensions, Introversion, Body, Suspicion, were the most nearly independent clusters (as Fig. 4 shows); only four pools of small residuals remained in the matrices of the four D, R, A, and T clusters.

Since the last of these, T Tension, had the greatest generality of the remaining four, it was decided to add T to I, B, and S as the final set of basic four dimensions of the MMPI.

Dimensional analysis of the 118 item-variables in the Patient Group.--A full cycle key cluster solution of the intercorrelations between the 118 items in the Patient group resulted in the cluster structure depicted in Fig. 4, top diagram. This factoring process was "preset" on the four basic dimensions defined by the items of I, B, S, and T. In the tight cluster at lower left in the configuration the symbols plotted as "I" and enclosed in a broken line are 15 of the 17 Introversion items that define this cluster. The remaining two lie nearby in the direction of the two arrows. In another tight cluster over at lower right are 16 Body, or B, items; the 17th item was dropped from the analysis because of trivial communality ( $h^2 < .1$ ). At the top you will find the Suspicion cluster. The remaining four clusters, Depression, Resentment, Autism, and Tension lie within the framework of the three I, B, S clusters. Clearly the total configuration for the Patient group shows an excellent cluster structure; it is virtually the same as that found previously in the total Inclusive group (Tryon, 1966b, Fig. 1).

Dimensional analysis of the 110 item-variables in the Normal Group.--A radically different dimensional structure emerges in the Normals, shown in the SPAN diagram of Fig. 4, lower. The dramatic change is in the Body cluster which was so sharply evident in the Patient group. It is absent as a distinct cluster among Normals! And so are the Depression or Autism clusters. But Introversion and Suspicion do appear as fairly independent item groups. Tension and Resentment also remain but move into a grand arc bounded by Introversion and Suspicion. It appears as if only Introversion and Suspicion are the dominant and distinctive dimensions of Normals in the MMPI item-clusters.

Comparison of the dimensions within each group separately (COMP 1).--Precise numerical statements about the seven item-clusters in each of the two groups are given in Table 6, section A

Fig. 4  
about  
here

Table 6  
about  
here

(analogous to Table<sup>6</sup> 2 in the Holzinger Problem), and sections C and D. The relationships between the seven domains represented as dimensions (or "oblique factors") are given in section A by the inter-domain  $r_{CC}$  values; those for the Patients are above the lined-off diagonal, those for Normals below. Recall that these "correlations between oblique factors" are merely abstract metric descriptions of the complex relationships depicted in the SPAN diagram, and though they are more precise numerical statements compared to the verbal statements about the configuration, they are more difficult to conceptually organize. And they can be misleading. I must leave to the reader a detailed examination of this complex table of relationships, suggesting that he cross-reference his study of it by simultaneously referring to the visual configuration in Fig.<sup>6</sup> 4.

Several obvious points may, however, be mentioned here. In both groups the Introversion and Suspicion dimensions are the most independent, and Tension is most positively correlated with all the other dimensions. But the Body dimension is radically different in the two groups, fairly specific in the Patients but rather general in the Normals, indeed correlating .90 with Autism! But this generality of the Body dimension is misleading in the Normal group, because from the configuration we know that Body is not a cluster-defined dimension in Normals but a mere sampling of heterogeneous items from their whole sphere of items. It is an omnibus grab-bag of items in the Normal, just as is Autism, so their high correlation is merely due to both being similar hodgepodes.

Direct comparative analysis of the dimensions across groups (COMP 2).--When we project the dimensions of the two groups into the same COMP2 analysis, we see directly and clearly the relations among the dimensions both within but especially across the two groups. They are pictorially displayed in the single SPAN diagram of Fig.<sup>6</sup> 5 (analogous to Fig.<sup>6</sup> 2 of the Holzinger Problem). The sharply differentiated and spread out dimensions of the Patients, denoted by the subscript "P" attached to the seven dimensions, I, B, S, D, R, A, T, confirms the within-group cluster structure of their items as previously depicted in the upper sphere of Fig.<sup>6</sup> 4. In contrast,

Fig.<sup>6</sup> 5  
about  
here



the within-group structure of the dimensions of the Normals, indicated by the subscript "N", confirms the narrow, essentially two-dimensional band ranging from Introversion to Suspicion.

Consider, now, the similarity of the dimensions across the two groups as objectively measured by the cos  $\theta$  values, given in Table<sup>6</sup>6, section B, especially those down the upper left to lower right diagonal. The most similar dimensions across the groups are Introversion (.71), Suspicion (.74), Resentment (.80), and Tension (.73). The least similar is Body (.49), a different kind of dimension in the two groups.

Attention is again drawn to the correspondence between the index of dimensional similarity, cos  $\theta$ , and the inter-domain ("common factor") correlations,  $r_{CC}$ , given as paired values in section A of Table<sup>6</sup>6 (analogous to section A of Table<sup>6</sup>2 in the Holzinger study). They show a close correspondence only for tight clusters I, S, R and T. Thus it is that in the comparative dimensional analysis of variables, the COMP 2 analysis accurately reveals the degree of similarity only of those dimensions defined by tight (highly collinear) clusters, a matter developed in technical Appendix A.

#### Comparative typological analysis in the MMPI Problem (OCOMP analysis).

The comparative typological objective is to discover the degree to which O-types of individuals, formed by classifying together individuals having the same pattern of Z-scores on the four basic MMPI dimensions, I, B, S, T, have the same structure in the Patient and Normal Groups. In this analysis, each person was scored by his Full Form scores on I, B, S, and T.

Similarity of frequency-patterns of the two groups on the common typology of the Inclusive Group.--In the typological analysis of the Inclusive group by program OTYPE (Iteration I<sub>1</sub>). I reported previously that 14 O-types emerged (Tryon, 1967a). These are listed as types M1 to M14 in Table<sup>6</sup>7, under "Inclusive typology", where you will find their frequencies, Z-scores on I, B, S, T, and descriptive names. When the Normal and Patient subjects are sorted to these 14 Inclusive O-types, the percentages of cases falling into them are

Table<sup>6</sup>7  
about  
here

the values listed in the "% in each group" columns. As a point of special interest, I separated the Patient group into its two component diagnostic groups, Anxieties and Schizophrenics.

The overall similarity of the typology of the three groups in relation to each other is given by their  $P$  values at the foot of the table (analogous to the presentation in the Holzinger Problem given in Table<sup>6</sup>3). Normals vs. Anxieties show a  $P = .17$ , indicating virtually no similarity in their typological structures. Curiously, there is a mild typological similarity of Normals and Schizophrenics, whose  $P = .41$ . The typologies of the Anxieties and Schizophrenics, in contrast, bear considerable resemblance, having a  $P = .73$ .

But the details of their differences, given in the column headed "Differences", are of great interest. Note that the Normals are almost exclusively concentrated in types M1, M2, and M3, described generally as Extrovert, Healthy, and Relaxed, with a few in M8, the Suspicious. The Anxieties excel in the Somatic types, M7, M9, M10, M13, M14, thus being persons most preoccupied by body disturbances. The Schizophrenics, compared to the Anxieties, behave typologically somewhat like Normals, excepting that they fall heavily in the Introvert type, M11.

#### Similarity of empirically-derived typologies of the groups.--

Fuller information on the differences between the O-types of the Normal and Patient Groups comes from<sup>a</sup> direct comparison of their typologies as these are empirically derived separately by the OTYPE and OSTAT programs but projected then into the same comparative analysis. In Table<sup>6</sup>8, left, you will find that, when the typology of the Normals is worked out independently, they fall into 14 types, N1 to N14, with no Unique individuals. In the right sector of the table, you will discover that the Patients were allocated to 12 O-types, P1 to P13.

As you look through the descriptive names of the Normal and Patient O-types, you may be astonished to discover that there is no overlap of their 27 types except for the Average and the Trusting O-types, but that even in these the Normals have only a handful of cases whereas they are abundant in the Patient group.

In sum, one finds that Patients are clearly distinguished

Table<sup>6</sup>8  
about  
here



from Normals in their objectively-derived patterns of MMPI scores. This finding goes directly to the question of the validity of the MMPI in distinguishing patients from normal persons. Our finding here definitely demonstrates the validity of the MMPI items in differentiating Normals from Patients, provided the item-cluster scores on I, B, S and T are used (and not the hodgepodge in the usual unclustered scales) and provided the objective typology described in these pages is used as the classificatory scheme.

When the 27 types are projected into the same EUCO-analysis (see the treatment of EUCO-analysis in the Holzinger Problem) along with the 14 Inclusive O-types, the grossly different typological structure of the Normals and Patients stands out boldly. This fact is clearly evident in the spherical representation of the types given in Fig.<sup>6</sup>6 (analogous to Fig.<sup>6</sup>3 of the Holzinger Problem). The Normal types, symbolized by "N" and placed in circles, are virtually all located in a super-cluster at the left in the "LOW" score ranges on all dimensions. The Patient types, symbolized by "P" placed in squares, are largely in the super-cluster at the right or "HIGH" region of the configuration. This separation confirms, of course, the finding of the previous section, but the SPAN configuration provides a more differentiated description.

Finally, observe the locus of the Inclusive types, symbolized by "M" and underlined. You will discover that these 14 types are located in all regions of this typological space where there are Normal and Patient types. This fact means that as a system of classifying individuals, normal or mentally-ill, the 14 Inclusive types, expounded on in more detail in an earlier paper (Tryon, 1967a), satisfactorily cover the ground.

#### Appendix A. Logic of $\cos \theta$ as an index of similarity between any two dimensions.

We begin by noting that within a group the index of similarity of any two dimensions,  $C_i$  and  $C_j$ , is the inter-domain correlation,

$$(1) \quad r_{C_i C_j} = \Sigma r_{ij} / \sqrt{\Sigma r_{ii}} \sqrt{\Sigma r_{jj}},$$

Fig.<sup>6</sup>6  
about  
here

where  $Zr_{1j}$  is the sum over the matrix of raw rs across definers of the two dimensions, and  $Zr_{11}$  and  $Zr_{jj}$  are sums over the raw rs within definers of each dimension. This is the old "correlation of sums" formula (Tryon, 1959, equation 24).

But this formula cannot be used in computing dimensional similarity across different groups since there are no raw rs between variables in different groups. But we do have the oblique factor coefficients of the n variables on dimensions  $C_1$  and  $C_j$  in different groups. Adjoining the matrices of factor coefficients of the two (or more) groups, we can compute the index of proportionality,  $P_{1j}$ , between factor coefficients of all pairs of dimensions within and across groups. This index is (Burt, 1948; Tucker, 1951; Wrigley & Neuhaus, 1955; Tryon, 1959):

$$(2) \quad P_{1j} = Zr_{vC_1} r_{vC_j} / \sqrt{Zr_{vC_1}^2} \sqrt{Zr_{vC_j}^2},$$

where  $r_{vC_1}$  and  $r_{vC_j}$  are the vectors of <sup>oblique</sup> factor coefficients of  $C_1$  and  $C_j$ . With a little algebra it can be shown that when the definers of any dimension have raw correlations that are perfectly collinear (are of rank 1), then we can solve for  $r_{C_1C_j}$  within a group using only the value of  $P_{1j}$ . The equation is (Tryon, 1962):

$$(3) \quad r_{C_1C_j} = \frac{1 - \sqrt{1 - P_{1j}^2}}{P_{1j}} = \cos \theta.$$

Expression (3) is called cos  $\theta$  because its magnitude is the cosine of the central angle between  $C_1$  and  $C_j$  when these dimensions are expressed as points on the hypersphere (the SPAN diagrams), such as that of Figs. 1 and 2, that is, whether or not they are dimensions within a group or across groups. The value of cos  $\theta$  gives exactly

the value of  $r_{CC}$  only when (1) the matrix of correlations between the definers of each dimension are of rank 1 and (2) when the adjoined vectors of their factor coefficients include as rows only the defining variables of the two dimensions. Otherwise,  $\cos \theta$  is only an approximation to  $r_{CC}$ . Program COMP2 of the BC TRY System computes  $\cos \theta$  for condition (2).

#### References

- Anastasi, A. Psychological testing. New York: MacMillan, 1961.
- Burt, C. The factorial study of temperamental traits. Brit. J. Psychol. Stat., 1948, 1, 178-203.
- Harman, H. Modern factor analysis. Chicago: Univ. of Chi. Press, 1960.
- Holzinger, K. and Swineford, F. A study in factor analysis: the stability of a bi-factor solution. Suppl. Educ. Monogr., No. 48, Chicago: Univ. of Chi. Press, 1939.
- Tryon, R. A theory of psychological components: an alternative to mathematical factors. Psych. Rev. 1935, 42, 425-454.
- Tryon, R. Cluster analysis. Ann Arbor, Michigan: Edwards Bros., 1939.
- Tryon, R. Reliability and behavior domain validity: Reformulation and historical critique. Psychometr., 1957, 54, 229-249.
- Tryon, R. Domain sampling formulation of cluster and factor analysis. Psychometr., 1959, 24, 113-135.
- Tryon, R. Theory of the BC TRY System: Statistical theory. Library version, ditto, 1964.
- Tryon, R. Identification of social areas by cluster analysis. Univ. Calif. Publ. Psychol., 8, No. 1, 1-100. Berkeley: U. Calif. Press

- Tryon, R. and Bailey, D. The BC TRY computer system of cluster and factor analysis. Mult. Behav. Res., 1966(a), 1, 95-111.
- Tryon, R. Unrestricted cluster and factor analysis with applications to the MMPI and Holzinger-Harman problems. Mult. Behav. Res., 1966(b), 1, 229-244.
- Tryon, R. Person-clusters on intellectual abilities and MMPI attributes. Mult. Behav. Res., (In press, 1967(a)).
- Tryon, R. Predicting individual differences in cluster analysis: Holzinger abilities and MMPI attributes. (Ms submitted for publication), 1967(b).
- Tryon, R. Predicting group differences in cluster analysis: The social area problem. (Ms submitted for publication), 1967(c).
- Tucker, L. A method of synthesis for factor analysis studies. Pers. Res. Sect. Rep., No. 984. Washington: Dept. Army, 1951.
- Wrigley, C. and Neuhaus, J. The matching of two sets of factors. Amer. Psychol., 1955, 10, 418-419 (Abstract).

#### Footnote

- <sup>1</sup> Supported in part by NSPE grants MH0811-01 to 04 and MH18134-01 to 05.

Table 6.1 The 24 Variables of the Holzinger Problem

## Spatial Tests

F1 VIS <sup>f</sup>Visual Figure Completions  
 F2 CUB <sup>f</sup>Cube Similarities  
 F3 PBD <sup>f</sup>Paper Form Board  
 F4 LOZ <sup>f</sup>Lozenge Shape Rotations

## Speed Tests

S10 ADD <sup>s</sup>Addition  
 S11 COD <sup>s</sup>Code Substitution  
 S12 CNT <sup>s</sup>Counting Groups of Dots  
 S13 SCC <sup>s</sup>Straight or Curved  
                   Capitals Discrimination

## Verbal Tests

V5 INF <sup>v</sup>General Information  
 V6 CMP <sup>v</sup>Paragraph Comprehension  
 V7 SNT <sup>v</sup>Sentence Completion  
 V8 WCL Word Classification  
 V9 WMN <sup>v</sup>Word Meaning (Vocabulary)

## Memory Tests

M14 WRG <sup>m</sup>Word Recognition  
 M15 NRG <sup>m</sup>Number Recognition  
 M16 FRG <sup>m</sup>Figure Recognition  
 M17 WN <sup>m</sup>Object-Number Recall  
 M18 NF <sup>m</sup>Number-Figure Recall  
 M19 FW Figure Word Recall

## Mathematical-Ability Tests

N20 DED Deduction  
 N21 PUZ Numerical Puzzles  
 N22 RSN Problem Reasoning  
 N23 SER Series Completion  
 N24 ARI Woody-McCall Mixed  
                   Fundamentals, Form I

- 
- <sup>f</sup> A definer of F(Space)  
<sup>v</sup> A definer of V(Verbal)  
<sup>s</sup> A definer of S(Speed)  
<sup>m</sup> A definer of M(Memory)

Table 6.2 Similarity of the Four Basic Holzinger Abilities, V, S, F, M, Within and Between the Suburban and Factory Groups.

$r_{CC}$  is the inter-domain  $r$ , "correlation between oblique factors", from the correlation of sums of  $r_s$ .<sup>a</sup>

$\cos \theta$  is the estimated  $r_{CC}$  from the index of proportionality,  $P$ , of the factor coefficients<sup>b</sup>

A. Similarity of cluster dimensions within each group<sup>a, b</sup>

		V Verbal	S Speed	F Form(Space)	M Memory
		$r_{CC}$ $\cos \theta$	$r_{CC}$ $\cos \theta$	$r_{CC}$ $\cos \theta$	$r_{CC}$ $\cos \theta$
V	Suburban	Unities	.43 .43	.58 .58	.46 .47
Verbal	Factory		.42 .43	.35 .37	.14 .14
S	Suburban	.43 .43	Unities	.53 .51	.56 .54
Speed	Factory	.42 .43		.29 .28	.39 .36
F	Suburban	.58 .58	.53 .51	Unities	.60 .56
Form (Space)	Factory	.35 .37	.29 .28		.27 .26
M	Suburban	.46 .47	.56 .54	.60 .56	Unities
Memory	Factory	.14 .14	.39 .36	.27 .26	

B. Similarity of cluster dimensions between groups ( $\cos \theta$  only)<sup>b</sup>

		Suburban			
		V <sub>s</sub>	S <sub>s</sub>	F <sub>s</sub>	M <sub>s</sub>
		$r_s$	$r_s$	$r_s$	$r_s$
Factory	V <sub>f</sub> Verbal	<u>.96</u>	.39	.48	.32
	S <sub>f</sub> Speed	.46	<u>.89</u>	.42	.48
	F <sub>f</sub> Form (Space)	.46	.36	<u>.92</u>	.39
	M <sub>f</sub> Memory	.28	.42	.41	<u>.83</u>

C. Generality of each dimension (reproducibility of  $r_s$ ).<sup>c</sup>

	Suburban			
	Factory	.51	.37	.47
		.50	.27	.28
				.40
				.18

D. Reliability coefficient ( $\alpha$ ) of cluster score on each dimension<sup>c</sup>

	Suburban			
	Factory	.90	.83	.70
		.90	.74	.69
				.76
				.73

<sup>a</sup>  $r_{CC}$  from CSA

<sup>b</sup>  $\cos \theta$  from COMPl

<sup>c</sup> From CSA

Table 6.3 Similarity of Frequency Patterns of Factory vs Suburban Children on the Common Inclusive Typology in the Holzinger Problem.

Inclusive typology <sup>a</sup>						Factory vs. Suburban			Boys vs Girls			
Types	Freq	Z-scores				Descriptive name	% in each group		Diff	% in each group		Diff
		V	S	F	M		P <sub>f</sub>	P <sub>s</sub>		P <sub>b</sub>	P <sub>g</sub>	
H1	14	48	36	44	37	Low Speed & Memory	4	6	-2	8	2	6(B)
H2	8	36	47	49	35	Low Verbal & Memory	5	0	5(F)	3	2	1
H3	21	48	50	48	38	Low Memory	4	10	-6(S)	10	5	5(B)
H4	9	50	39	36	48	Low Speed & Form	2	4	-2	2	4	-2
H5	13	36	42	36	45	Low Verbal & Form	5	4	1	2	6	-4(G)
H6	20	50	50	38	46	Low Form	6	7	-1	5	8	-3(G)
H7	19	47	37	48	50	Low Speed	3	10	-7(S)	8	5	3(B)
H8	23	35	47	53	54	Low Verbal	13	1	12(F)	9	6	3(B)
H9	38	51	51	48	51	Average	13	12	1	10	15	-5(G)
H10	22	65	50	53	47	H1 Verbal	3	12	-9(S)	7	8	-1
H11	23	47	65	51	52	H1 Speed	11	4	7(F)	6	9	-3(G)
H12	14	64	64	59	58	H1 Verbal & Speed	3	6	-3(S)	3	6	-3(G)
H13	27	52	51	63	49	H1 Form	9	9	0	13	5	8(B)
H14	23	52	49	51	63	H1 Memory	9	6	3(F)	5	10	-5(G)
H15	8	57	63	54	67	H1 Speed & Memory	3	3	0	2	3	-1
Unique	19					Unique	7	6	1	7	6	1
N	301						100	100		100	100	
						N	156	145		146	155	

<sup>a</sup>Iteration I<sub>2</sub> of OTYPE

From OTYPE and OSTAT

Similarity of frequency patterns of Factory and Suburban children

$$P_{fs} = \sum p_f p_s / \sqrt{\sum p_f^2} \sqrt{\sum p_s^2} = 617 / \sqrt{828} \sqrt{820} = .75$$

Similarity of frequency patterns of Boys and Girls

$$P_{bg} = 667 / \sqrt{792} \sqrt{786} = .85$$

Significance

For  $m = 16$  types, the mean proportion in them is  $p' = .06$ , whence:

$$3\sigma_d = 3(10\sqrt{p'q'(1/N_f + 1/N_g)}) = 30\sqrt{(.06)(.94)(1/156 + 1/145)} = 3.$$

Table 6.4 Within-group Typologies of the Factory and Suburban Children in the Holzinger Problem.

Factory Children							Suburban Children								
Type	Freq	Profile level and homogeneity					Descriptive name	Type	Freq	Profile level and homogeneity					Descriptive name
		V	S	F	M	H				V	S	F	M	H	
F1	7	<u>34</u>	<u>36</u>	46	<u>37</u>	.81	Low Verbal, Speed & Mem	S1	8	51	<u>35</u>	48	<u>36</u>	.86	Low Speed & Memory
F2	5	<u>40</u>	52	52	<u>34</u>	.87	Low Verbal & Memory	S2	13	48	51	43	<u>37</u>	.92	Low Memory
F3	14	57	50	57	<u>40</u>	.86	Low Memory	S3	14	50	43	<u>36</u>	45	.87	Low Form
F4	18	41	45	<u>32</u>	45	.87	Low Form	S4	17	45	<u>36</u>	46	49	.86	Low Speed
F5	5	48	54	<u>40</u>	44	.91	Low Form	S5	4	<u>36</u>	48	44	50	.94	Low Verbal
F6	9	53	46	45	46	.89		S6	21	51	51	49	49	.93	
F7	18	42	57	47	51	.87		S7	16	<u>66</u>	51	53	49	.92	H1 Verbal
F8	11	58	55	51	55	.87		S8	10	52	<u>62</u>	49	53	.86	H1 Speed
F9	4	47	<u>66</u>	48	46	.83	H1 Speed	S9	6	<u>66</u>	<u>67</u>	59	57	.35	H1 Verbal & Speed
F10	4	52	<u>70</u>	55	58	.91	H1 Speed	S10	15	52	49	<u>63</u>	50	.86	H1 Form
F11	15	40	48	<u>62</u>	51	.77	H1 Form	S11	12	56	53	52	<u>64</u>	.85	H1 Memory
F12	5	<u>63</u>	<u>63</u>	<u>67</u>	53	.75	H1 Verbal, Speed & Form	S12	3	<u>68</u>	50	<u>62</u>	<u>69</u>	.94	H1 Verbal, Form & Memory
F13	23	43	50	53	59	.82		Uniq	6						
F14	9	52	51	48	<u>66</u>	.79	H1 Memory								
Uniq	0														
N	156							N	145						
Eta		.85	.81	.84	.83			Eta		.87	.91	.87	.90		

From OSTAT

\* Iteration 1<sub>2</sub> of OTYPE



Table 6.5 Defining Items of the Seven Item-Clusters of the MMPI

## A. The four basic item-clusters

I: Introversion

(Full Form, 26 items, rel. .93; Short Form, 1st 17 items, rel. .91)

377	180	86	52	292	138	-415
- 57	-371	171	-309	- 79	-353	-482
321	267	-547	-479	317	304	
201	172	-521	509	-264	-449	

B: Body symptoms

(Full Form, 33 items, rel. .92; Short Form, 1st 17 items, rel. .89)

-243	62	47	125	161	- 36	-160	-330	14
189	-175	44	- 68	544	-163	191	- 2	
108	-230	- 55	10	72	- 51	-153	- 18	
-190	114	29	23	- 3	-103	263	-192	

S: Suspicion and mistrust

(Full Form, 25 items, rel. .85; Short Form, 1st 17 items, rel. .83)

404	436	368	447	406	89	455
507	136	280	319	278	112	
383	244	265	71	284	426	
390	348	469	558	438	316	

T: Tension, worry and fears

(Full Form, 36 items, rel. .92; Short Form, 1st 17 items, rel. .88)

555	238	43	448	338	439	158	322	-131
431	506	-242	186	-407	335	303	360	365
337	543	340	499	182	102	13	22	494
217	442	-152	166	32	473	388	351	492

## B. The three remaining "dependent" item-clusters

D: Depression and apathy

(Full Form, 28 items, rel. .94; Short Form, 1st 17 items, rel. .91)

76	-379	418	424	142	397	84	- 88
-107	487	- 8	396	526	357	- 46	
236	41	549	61	361	168	104	
301	259	67	411	384	339		

R: Resentment and aggression

(Full Form, 21 items, rel. .87; Short Form, 1st 16 items, rel. .82)

94	375	536	145	416	443
336	39	139	148	382	
468	381	234	28	106	
-399	97	129	162	147	

A: Autism and disruptive thoughts

(Full Form, 23 items, rel. .86; Short Form, 1st 17 items, rel. .81)

559	425	560	342	33	40
241	511	-329	374	359	31
15	545	100	459	389	134
349	358	345	297	356	

Table 6.6 Similarity of the Seven MMPI Item-Cluster Dimensions Within and Between the Normal and Patient Groups.

$r_{CC}$  is the inter-domain  $r$ , "correlation between oblique factors", from the correlation of sums of  $r_s$  <sup>a</sup>

$\text{Cos } \theta$  is the estimated  $r_{CC}$  from the index of proportionality,  $\frac{r_{CC}}{P_i}$  of the factor coefficients <sup>b</sup>

A. Similarity of item-cluster dimensions within each group <sup>a, b</sup>

	I Introversion		B Body		S Suspicion		D Depression		R Resentment		A Autism		T Tension	
	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$	$r_{CC}$	$\text{Cos } \theta$
I Introversion		P	.12	.13	.31	.31	.71	.69	.47	.46	.33	.38	.50	.49
B Body	.62	.46		A	.34	.34	.32	.31	.37	.37	.50	.49	.63	.60
S Suspicion	.07	.06	.61	.52		R	.38	.37	.66	.62	.65	.61	.59	.57
D Depression	.76	.60	.56	.43	.43	.38		M	.66	.65	.57	.55	.78	.76
R Resentment	.37	.32	.64	.52	.76	.69	.76	.65		A	.64	.62	.79	.77
A Autism	.51	.45	.90	.69	.77	.70	.72	.59	.74	.68		L	.77	.74
T Tension	.59	.53	.72	.53	.50	.48	.72	.59	.71	.63	.67	.61		S

B. Similarity of item-cluster dimensions between groups ( $\text{Cos } \theta$  only) <sup>b</sup>

		Patients						
		I <sub>P</sub>	B <sub>P</sub>	S <sub>P</sub>	D <sub>P</sub>	R <sub>P</sub>	A <sub>P</sub>	T <sub>P</sub>
N O R M A L	I <sub>N</sub> Introversion	.71	.20	.22	.53	.42	.38	.49
	B <sub>N</sub> Body	.33	.49	.47	.32	.49	.51	.50
	S <sub>N</sub> Suspicion	.15	.30	.74	.28	.60	.57	.46
	D <sub>N</sub> Depression	.56	.31	.43	.61	.65	.58	.59
	R <sub>N</sub> Resentment	.33	.31	.64	.48	.80	.57	.62
	A <sub>N</sub> Autism	.35	.42	.61	.45	.60	.69	.60
	T <sub>N</sub> Tension	.48	.41	.53	.52	.70	.57	.73

C. Generality of each dimension (reproducibility of  $r_s$ ) <sup>c</sup>

Normals	.38	.51	.43	.53	.52	.52	.48
Patients	.31	.19	.24	.52	.42	.36	.61

D. Reliability coefficient ( $\alpha$ ) of cluster score on each dimension <sup>c</sup>

Normals	.81	.54	.83	.72	.80	.76	.75
Patients	.90	.87	.83	.88	.79	.79	.81

<sup>a</sup>  $r_{CC}$  from CSA

<sup>b</sup>  $\text{Cos } \theta$  from COMP2

<sup>c</sup> From CSA

Table 6.7 Similarity of the Frequency Patterns of Normals and Patients on the Common Inclusive Typology in the MMPI Problem.

Inclusive typology						% in each group	Differences						
Type	Freq	Z-scores				Descriptive	Norm P <sub>N</sub>	Anx P <sub>A</sub>	Schiz P <sub>S</sub>	N-A	N-E	A-S	
		I	B	S	T								
M1	24	<u>40</u>	<u>38</u>	<u>35</u>	<u>35</u>	Extro-Healthy	20	1	6	19(N)	14(N)	-5(S)	
M2	38	<u>37</u>	<u>40</u>	48	<u>38</u>	Extro-Healthy -Trust-Relaxed	39	1	3	38(N)	36(N)	-2	
M3	21	46	<u>39</u>	48	<u>40</u>	Healthy-Relaxed	16	1	7	15(N)	9(N)	-6(S)	
M4	31	48	47	<u>36</u>	46	Trusting	5	15	6	-10(A)	-1	9(A)	
M5	17	<u>39</u>	50	50	47	Extrovert	3	5	10	-2	-7(S)	-5(S)	
M6	24	50	50	50	50	Average	3	11	7	-8(A)	-4(S)	4(A)	
M7	22	52	<u>62</u>	52	<u>64</u>	Somatic-Tense	0	11	7	-11(A)	-7(S)	4(A)	
M8	26	48	48	<u>64</u>	52	Suspicious	12	3	14	9(N)	-2	-11(S)	
M9	16	50	<u>65</u>	52	52	Somatic	0	10	1	-10(A)	-1	9(A)	
M10	14	50	<u>64</u>	<u>66</u>	<u>60</u>	Somatic-Suspicious -Tense	0	8	3	-8(A)	-3	5(A)	
M11	30	<u>64</u>	48	48	50	Introvert	2	10	19	-8(A)	-17(S)	-9(S)	
M12	20	<u>66</u>	54	<u>60</u>	<u>64</u>	Intro-Suspicious -Tense	0	9	10	-9(A)	-10(S)	-1	
M13	17	<u>65</u>	<u>65</u>	53	<u>66</u>	Intro-Somatic -Tense	0	9	4	-9(A)	-4(S)	5(A)	
M14	10	<u>67</u>	<u>67</u>	<u>65</u>	<u>67</u>	Intro-Somatic -Suspicious-Tense	0	5	3	-5(A)	-3	2	
Uniq	0					Unique	0	0	0	0	0	0	
N		310					N	90	150	70			

#### Similarity of frequency patterns

From OTYPE and OSTAT

##### Normals vs Anxiety

$$P_{na} = \sum p_n p_a / \sqrt{\sum p_n^2 \sum p_a^2} = 254 / \sqrt{2368 \cdot 955} = .17$$

##### Normals vs Schizo

$$P_{ns} = 636 / \sqrt{2368 \cdot 1,020} = .41$$

##### Anxiety vs Schizo

$$P_{as} = 717 / \sqrt{955 \cdot 1,020} = .73$$

#### Significance

For  $m = 15$  types, the mean proportion in them is  $p' = .07$ , whence (See Table 3):

$$\text{Normals vs Anxiety: } 3\sigma_d = 30\sqrt{(.07)(.93)(1/90 + 1/150)} = 3.2$$

$$\text{Normals vs Schizo: } 3\sigma_d = 30\sqrt{(.07)(.93)(1/90 + 1/70)} = 3.8$$

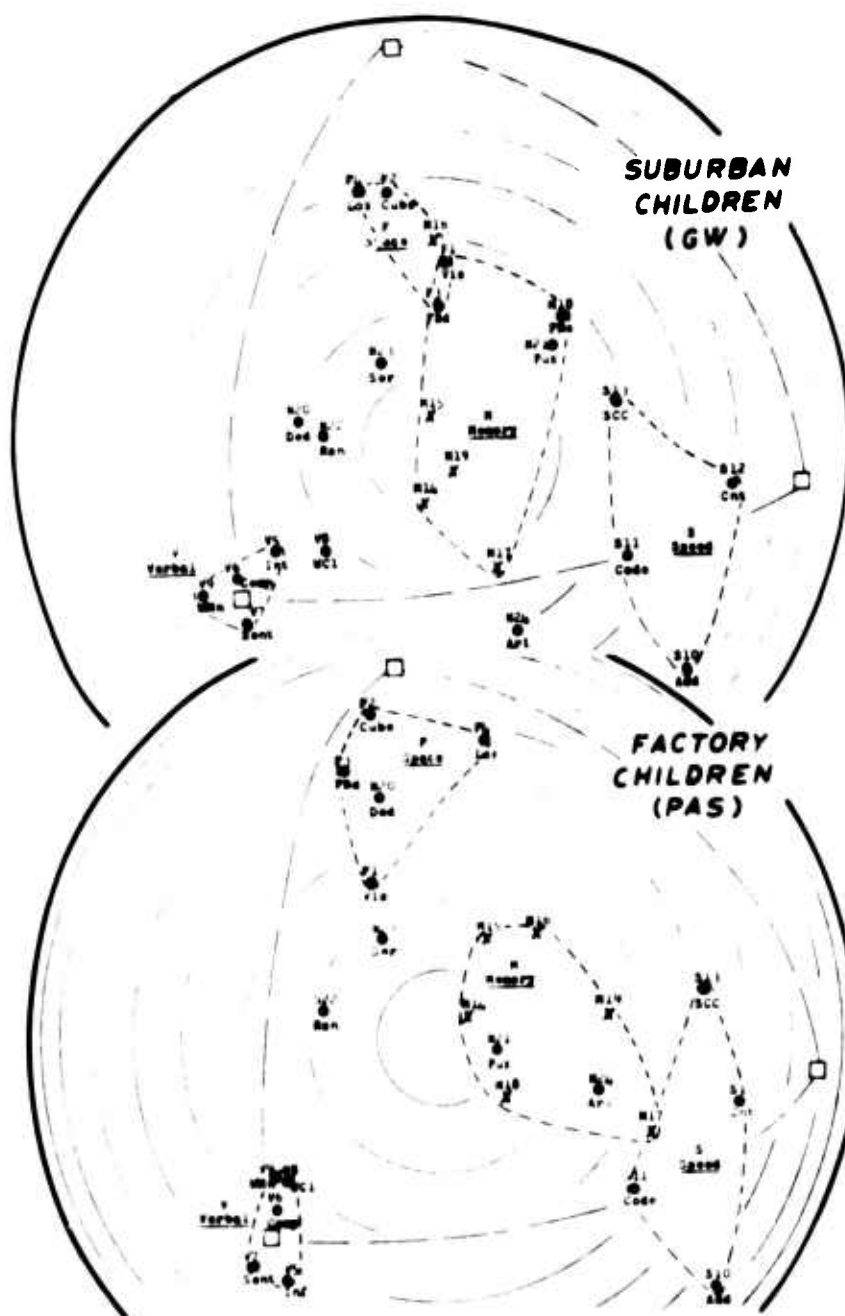
$$\text{Anxiety vs Schizo: } 3\sigma_d = 30\sqrt{(.07)(.93)(1/150 + 1/70)} = 3.5$$

Table 6.8 Within-group MMPI Item-cluster Typologies  
of the Normals and Patients

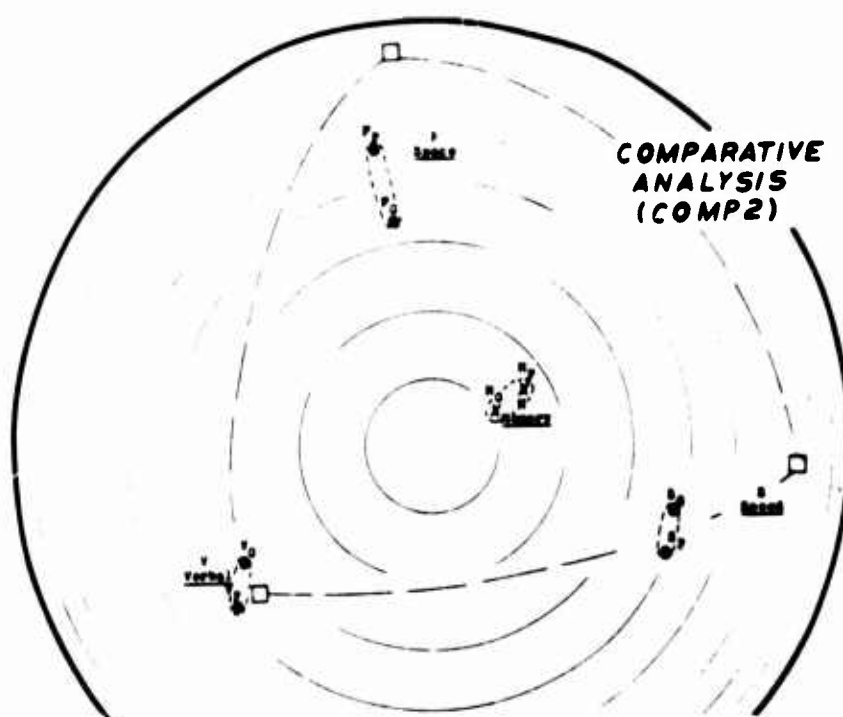
Normals							Patients								
Type	Freq	Profile level and homogeneity					Descriptive name	Type	Freq	Profile level and homogeneity					Descriptive name
		I	B	S	T	H				I	B	S	T	H	
N1	8	<u>38</u>	<u>39</u>	<u>34</u>	<u>35</u>	.99	Extro-Healthy -Trust-Relax	P1	36	<u>45</u>	<u>45</u>	<u>37</u>	<u>44</u>	.85	Trusting
N2	10	<u>41</u>	<u>39</u>	<u>38</u>	<u>37</u>	.98	Healthy-Trust -Relaxed	P2	15	<u>41</u>	<u>51</u>	<u>51</u>	<u>48</u>	.90	Average 1
N3	3	<u>38</u>	<u>42</u>	<u>37</u>	<u>37</u>	.99	Extro-Trust -Relaxed	P3	28	<u>49</u>	<u>47</u>	<u>53</u>	<u>50</u>	.89	Average 2
N4	16	<u>39</u>	<u>38</u>	<u>47</u>	<u>37</u>	.98	Extro-Healthy -Relaxed	P4	28	<u>62</u>	<u>47</u>	<u>44</u>	<u>51</u>	.88	Introvert
N5	3	<u>45</u>	<u>39</u>	<u>44</u>	<u>36</u>	.98	Healthy-Relax	P5	15	<u>49</u>	<u>63</u>	<u>46</u>	<u>50</u>	.84	Somatic
N6	4	<u>39</u>	<u>45</u>	<u>48</u>	<u>39</u>	.97	Extro-Relaxed	P6	10	<u>51</u>	<u>50</u>	<u>65</u>	<u>57</u>	.93	Suspicious
N7	4	<u>45</u>	<u>43</u>	<u>52</u>	<u>39</u>	.97	Relaxed	P7	12	<u>49</u>	<u>64</u>	<u>64</u>	<u>59</u>	.89	Somatic-Suspicio
N8	4	<u>53</u>	<u>41</u>	<u>38</u>	<u>41</u>	.95	Trusting	P8	15	<u>54</u>	<u>57</u>	<u>52</u>	<u>62</u>	.93	Tense
N9	6	<u>39</u>	<u>39</u>	<u>45</u>	<u>42</u>	.98	Extro-Healthy	P9	11	<u>64</u>	<u>50</u>	<u>57</u>	<u>60</u>	.90	Introvert-Tense
N10	3	<u>44</u>	<u>42</u>	<u>52</u>	<u>45</u>	.97	Average 1	P10	12	<u>49</u>	<u>66</u>	<u>49</u>	<u>64</u>	.90	Somatic-Tense
N11	12	<u>39</u>	<u>40</u>	<u>57</u>	<u>39</u>	.96	Extro-Healthy -Relaxed	P11	14	<u>65</u>	<u>67</u>	<u>50</u>	<u>64</u>	.90	Introvert -Somatic-Tense
N12	6	<u>53</u>	<u>45</u>	<u>52</u>	<u>48</u>	.91	Average 2	P12	14	<u>65</u>	<u>54</u>	<u>64</u>	<u>60</u>	.90	Intro-Suspicio -Tense
N13	5	<u>39</u>	<u>46</u>	<u>66</u>	<u>45</u>	.96	Extro-Suspicio	P13	10	<u>65</u>	<u>68</u>	<u>64</u>	<u>66</u>	.92	Intro-Somatic -Suspicio-Tense
N14	6	<u>44</u>	<u>47</u>	<u>60</u>	<u>46</u>	.98	Suspicious	Unique 0							
Uniq	0														
N	90							N	220						
Eta		.97	.98	.97	.97					.91	.88	.88	.90		

From OSTAT

<sup>a</sup> Iteration I<sub>2</sub> of OTYPE



**Fig.6.1 Cluster structure of abilities within the Suburban and Factory groups**



**Fig. 6.2 Cluster structure of abilities  
across the Suburban and Factory groups**

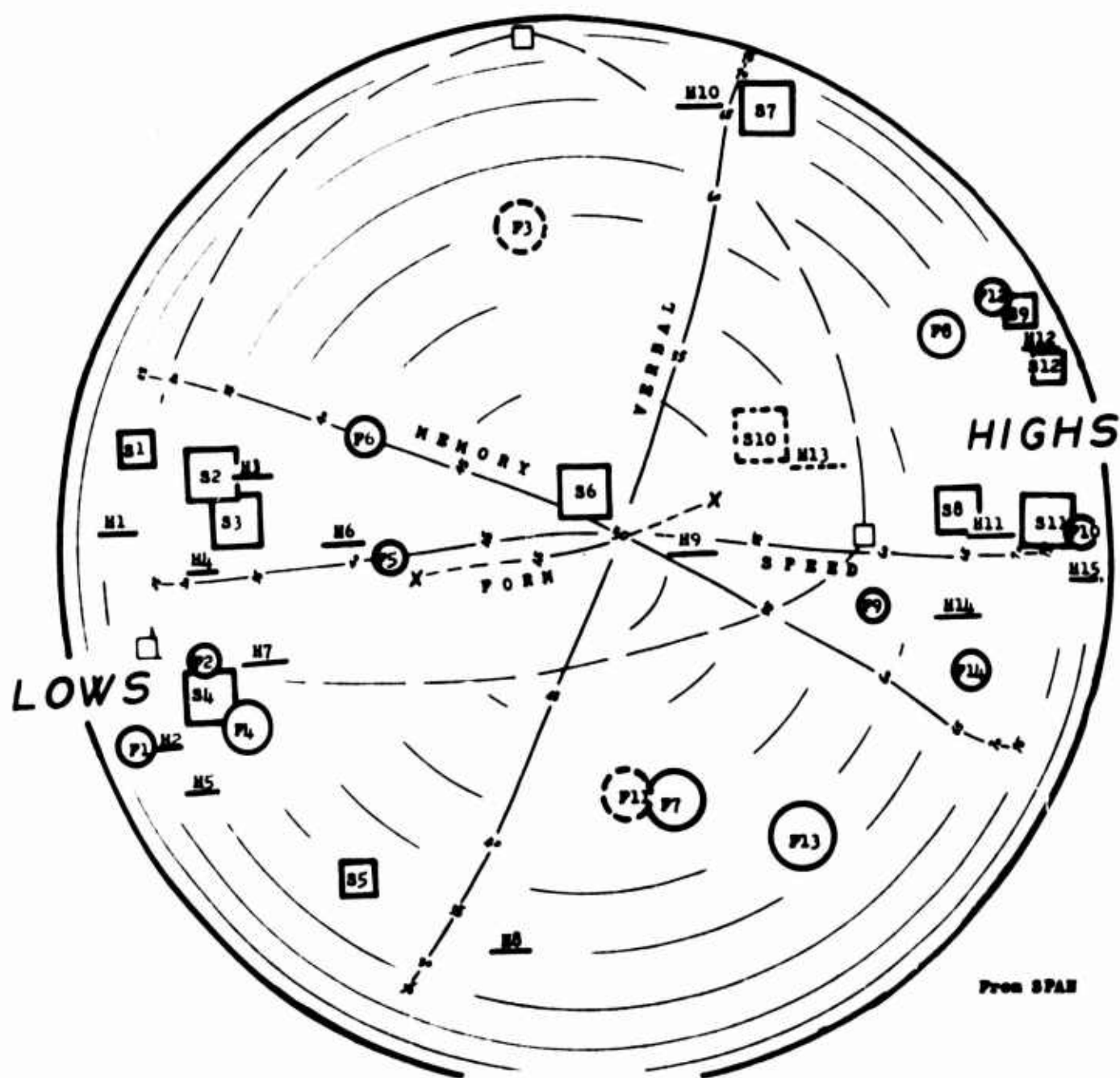


Fig. 6.3 Spherical representation of the typologies of the Pastory and Suburban Groups

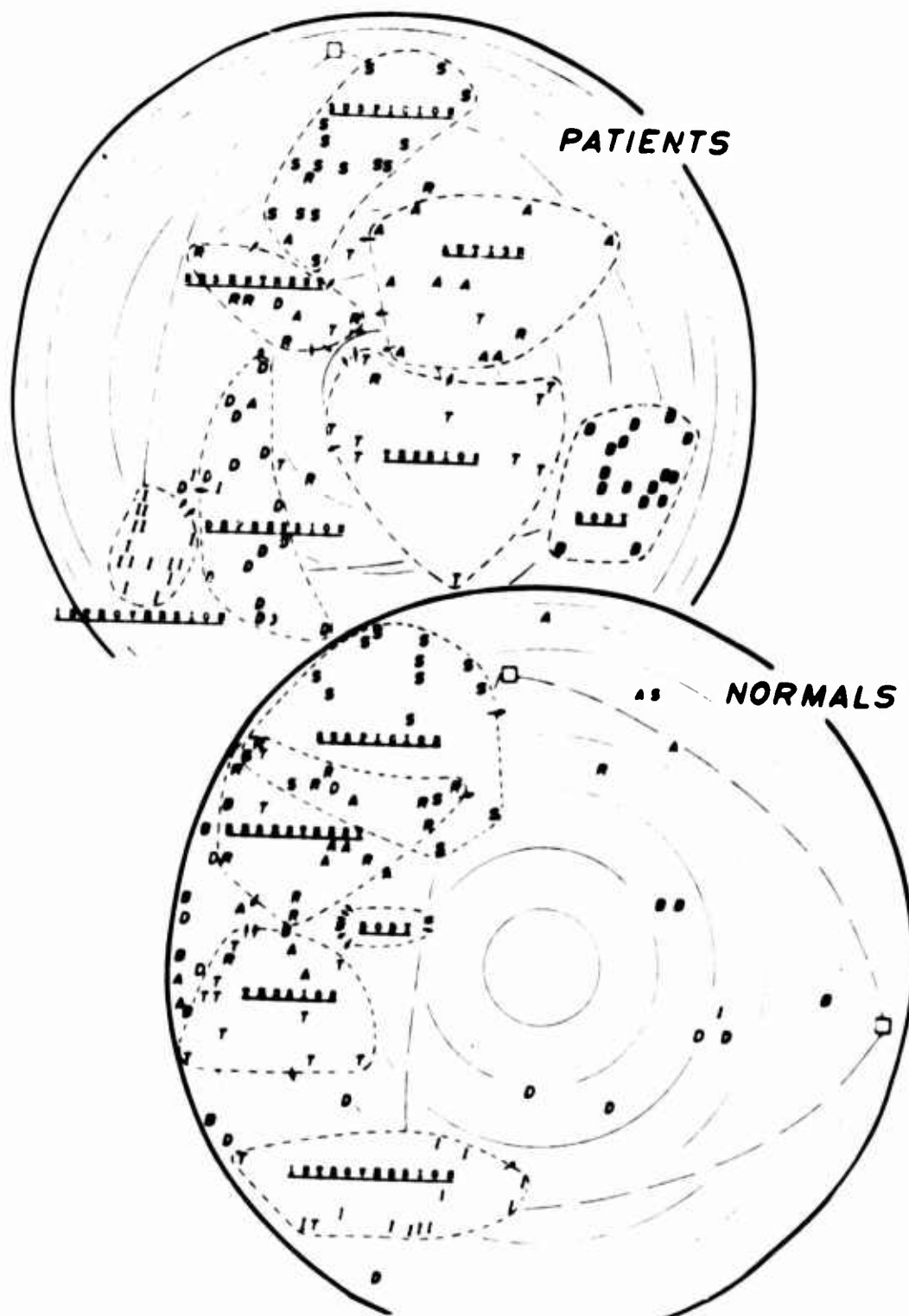


Fig. 6.4 Cluster structure of 118 MMPI items within the Patient and Normal groups



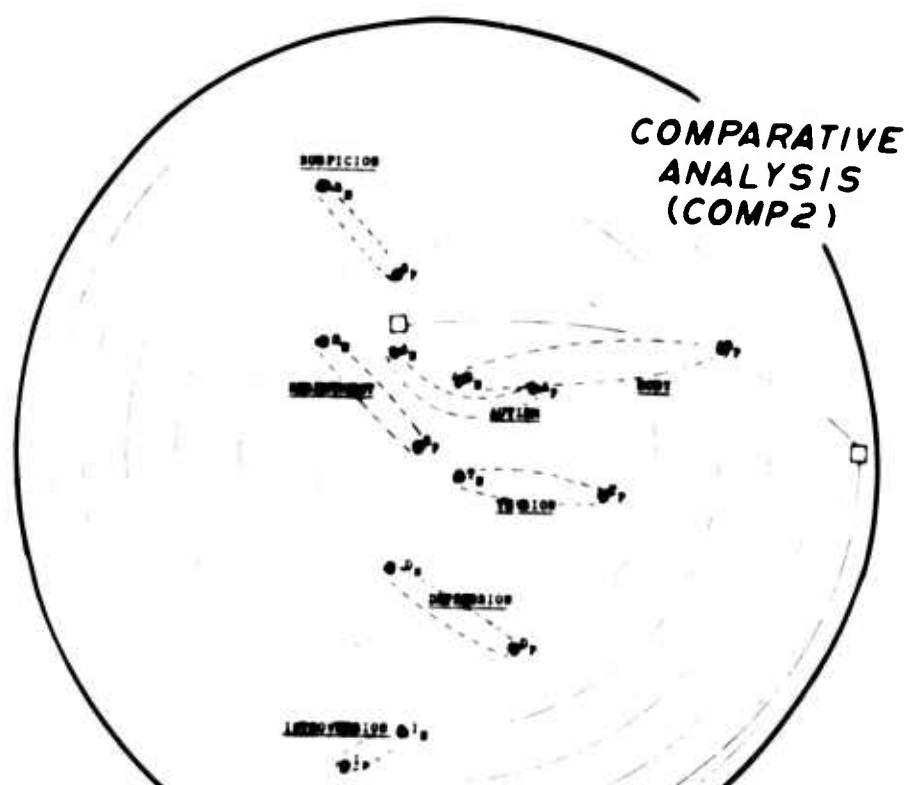


Fig.6.5 Cluster structure of 118 MMPI items across the Patient and Normal groups

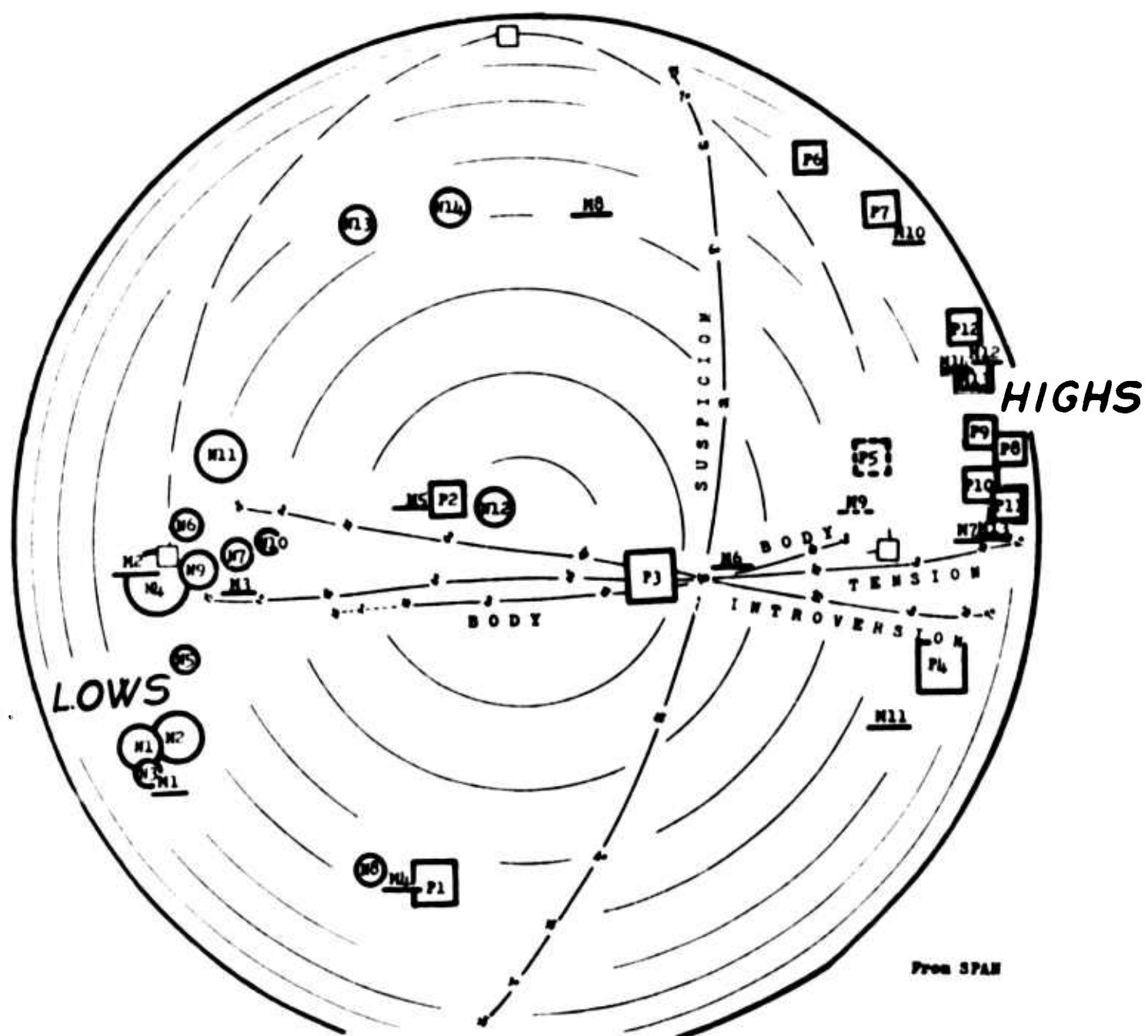


Fig. 6.6 Spherical representation of the typologies of the Normals and Patients

STANFORD RESEARCH INSTITUTE

MENLO PARK, CALIFORNIA



A COMPARISON OF TWO TECHNIQUES FOR FINDING THE MINIMUM  
SUM-SQUARED ERROR PARTITION

by Geoffrey H. Ball  
Senior Research Engineer  
Stanford Research Institute  
Menlo Park, California

I INTRODUCTION

Two difficult problems associated with cluster-seeking techniques are the comparison of cluster-seeking techniques and problems in interpreting the results of running cluster-seeking techniques on data.

In this paper, two techniques for finding minimum squared-error clusters are described and two recent results related to these techniques are discussed. Some sets of data for examination and evaluation of cluster-seeking techniques are given and the two techniques--ISODATA and the Singleton-Kautz algorithm--are compared. In addition, some useful graphical presentations for showing the structure of a body of data are presented. Methods that we have found helpful for interpreting experimental results are also discussed.

## II TECHNIQUES FOR FINDING MINIMUM SQUARED ERROR CLUSTERS

Before describing the Singleton-Kautz algorithm and the ISODATA algorithm, we discuss briefly the partitioning of a data set into minimum squared error (MSE) clusters. By partitioning we mean the assignment of each data point to one and only one of  $k$  subsets. In MSE partitions we wish to find the assignments of the data points to the  $k$  subsets or clusters that minimize the squared error. This squared error consists of the sum of the distances, taken over all the data points, from the data point to the point that lies at the mean of the cluster to which the data point is assigned.

A convenient representation of squared error is by using the sum of the products matrices  $T$ ,  $W$ , and  $B$ .<sup>\*</sup> The within-sum-of-products matrix  $W$  is a constant multiple of the pooled covariance matrix of data points. It is obtained by subtracting its associated cluster average point from each data sample and then calculating  $N$  times the covariance matrix for this reduced set of points, where  $N$  is the number of data samples. The between-sum-of-products matrix  $B$  gives the amount and direction of the deviation of the cluster centers from the overall mean, weighted by the number of data points in each cluster. The sum  $T = (W + B)$  is a constant matrix independent of the partitioning of the data points.<sup>\*\*</sup>

<sup>\*</sup> Formally,

$$W = [w_{ij}] \text{ where}$$

$$w_{ij} = \sum_{g=1}^G \sum_{k=1}^{N_g} (x_{gik} - \frac{1}{N_g} \sum_{l=1}^{N_g} x_{gil}) \cdot (x_{gjk} - \frac{1}{N_g} \sum_{l=1}^{N_g} x_{gjl})$$

where  $x_{gjl}$  is from the  $g^{\text{th}}$  group and is the  $j^{\text{th}}$  component of the  $l^{\text{th}}$  data point, and  $B = [b_{ij}]$  where

$$b_{ij} = \sum_{g=1}^G N_g (\frac{1}{N_g} \sum_{l=1}^{N_g} x_{gil} - \frac{1}{N} \sum_{m=1}^G \sum_{l=1}^{N_m} x_{mil}) (\frac{1}{N_g} \sum_{l=1}^{N_g} x_{gjl} - \frac{1}{N} \sum_{m=1}^G \sum_{l=1}^{N_m} x_{mjl}).$$

Other symbols used are:

$G$ , the number of clusters,  $N_g$  the number of data points in the  $g^{\text{th}}$  cluster, and  $N = \sum_{g=1}^G N_g$ , the total number of data points.

<sup>\*\*</sup> See Friedman and Rubin (1966) for a more detailed discussion of these matrices.

The eigenvalues and corresponding eigenvectors of  $W^{-1}B$  play a central role in discriminant analysis. Any function of the eigenvalues and the corresponding eigenvectors of  $W^{-1}B$  is invariant under linear transformations of the data. Useful functions of these eigenvalues are: the product of the eigenvalues, the maximum eigenvalue, and the sum of the eigenvalues.

Another simple function of these matrices is the sum of the diagonal elements. This can be represented symbolically by using the "trace" operator, which is a linear operator. From  $T = W + B$  we get  $\text{trace } T = \text{trace } W + \text{trace } B$ . The trace of a matrix also is the sum of the eigenvalues of that matrix.

From this equation we can see that for a given set of data, minimization of trace  $W$  results in the maximization of trace  $B$ . Trace  $T$ , as we have previously noted, is constant for a fixed data set with respect to modifications of the partitioning of the data set. Note that trace  $T$  is not invariant with respect to linear transformations. The MSE partition of a data set is the partition of the data set that minimizes trace  $W$ .

Another important function of the  $W$  matrix is the Mahalanobis type of distance, which can be written as

$$\frac{1}{N} (x-y)W^{-1}(x-y)',$$

where  $x$  is one point of the data set and  $y$  is another point. This distance is also invariant with respect to linear transformations of the data set. It is not invariant, of course, to new groupings of the data since, in general, this changes  $W^{-1}$ . The matrix  $W$  can be viewed as a linear transformation of the original data, since the Mahalanobis type distance between  $x$  and  $y$  can be rewritten as

$$(u - v) (u - v)' = (xW^{-\frac{1}{2}} - yW^{-\frac{1}{2}}) (xW^{-\frac{1}{2}} - yW^{-\frac{1}{2}})'$$

Note that this is the Euclidean distance between  $u$  and  $v$ , where  $u$  and  $v$  are obtained by linear transformation by  $W^{-\frac{1}{2}}$  from  $x$  and  $y$ , respectively. The difficulty lies in the necessity to compute  $W^{-1}$  for each partition of the data to be evaluated in the minimization, since  $W$  changes when the partition changes. We discuss below this minimization of the sum of the Mahalanobis type distances while simultaneously changing  $W$  as the partition is altered.

Dr. Richard Singleton has shown that for a MSE partition it is necessary (but not sufficient) that the hyperplane that is the perpendicular bisector of the line connecting any two cluster means cannot intersect the convex hulls of those two clusters. (This requires that the convex hull\* of one cluster not intersect the convex hull of another cluster.)

---

\* The minimum volume convex body sufficient to contain all of the data points in one cluster. If the volume is zero (i.e., the data is linearly dependent), then minimize the volume in the linear subspace of highest dimensionality in which the volume is non-zero.

It follows from this condition that a partition can be a stable MSE partition only if the means of the respective clusters are such that the above condition is satisfied. As we describe below, the ISODATA procedure uses this condition to seek an MSE partition by reassigning patterns that do not meet that condition to that cluster having the closest cluster center. The use of the perpendicular bisector can be generalized to distances measured using the Mahalanobis type distance.

#### The Singleton-Kautz Algorithm

The Singleton-Kautz algorithm was developed by Dr. Richard K. Singleton and Dr. William Kautz of Stanford Research Institute in 1965.\* This algorithm seeks explicitly to minimize trace W. The algorithm uses the following steps to perform this minimization.

- (1) All data points are assigned to a single partition.
- (2) The data point farthest from the single cluster mean is assigned to a second cluster.
- (3) All data points are sequentially tested to determine if a reassignment to the second cluster will reduce the sum-squared error (SSE). (Fortunately the computation requires only the evaluation of the change resulting from the reassignment.)\*\*
- (4) When it is no longer possible to reduce the SSE by reassigning any single data point to a different cluster, then the number of clusters is increased by one and the process is repeated. For data sets of 200 points experience indicates that about four cycles through the data point lead to the situation in which no single data point reassignment will result in a reduction of the SSE. For larger numbers of data points the number of cycles may increase considerably. (See discussion of this point in Sec. V, below.)

---

\*For similar techniques see also Forgy (1966) and Friedman & Rubin (1966).  
 \*\*The quantity calculated for the cluster to which data point  $x_i$  is assigned

is  $\frac{N_G}{N_G - 1} \sum_{i=1}^D (x_i - S_i^G / N_G)^2$  where  $N_G$  is the number of data points the  $G^{th}$

cluster;  $S_i^G$  is the sum of the  $i^{th}$  coordinate of all points in the  $G^{th}$  cluster excluding  $x_i$ . This quantity is compared with

$$\frac{N_H}{N_H + 1} \sum_{i=1}^D (x_i - S_i^H / N_H)^2 \text{ for clusters } H \text{ and the data point moved}$$

if the former exceeds the latter.

- (5) The number of clusters is increased to some maximum number of clusters. This maximum number is set by the person using the program.
- (6) When the limit is reached then the cycle is reversed and the number of clusters is changed by reducing the number of clusters by combining those two clusters that minimally increase the SSE. After combining those two clusters, the cycle described above is then used to attempt to further decrease the SSE. If the SSE found on this stage through the cycle is smaller than the SSE found in any previous stage, then the partition obtained on this new clustering is substituted for the partitioning found the previous time.
- (7) This increasing and decreasing of the number of clusters is continued until it is not possible to reduce the SSE further. At this point the process terminates.

Critical steps in this process are the selection of the data point used to initiate a new cluster, the ordering of data points, and the choice of those two clusters that are to become combined when the number of clusters is reduced. These comments can be summarized by saying that the choice of the starting points for the iterative hill-climbing to an MSE partition determines whether the partition obtained is the minimum among all local minimum squared error partitions of the data set.

#### The ISODATA Algorithm

The name ISODATA (see Ball and Hall, 1965, or Ball and Hall, 1966) applies to a variety of similar cluster-seeking techniques.\*\* The defining characteristics of these techniques are:

- (1) the iterative nature of the algorithm
- (2) the partitioning of all the data points into subsets without changing the cluster averages, such that data points are assigned to the closest previously obtained cluster average,
- (3) the combining of pairs of clusters into a single cluster,
- (4) the splitting of single clusters into a pair of clusters.

---

\* These data points could be, but at present are not, randomly reordered after each sequence of evaluations in order to reduce any sequential effects of taking the patterns one at a time in a particular order.

\*\* See also Sebestyen and Edie (1964), Sebestyen (1966), MacQueen (1966), and Stark (1962) for similar techniques.

Figure 7.1 shows a pictorial flow diagram of ISODATA. The patterns are sorted, one by one, on the basis of a measure of distance from a set of initial cluster points. Each pattern goes into that subset having the cluster point to which it is closest.

After all patterns have been sorted into one of the clusters the average of each of these subsets of patterns is computed and for each subset the standard deviation in each dimension are determined. These values are then passed into the Cluster Information Hopper.

The individual sample points in small clusters(those with fewer than  $\theta_N$  elements are considered small) are removed from the data set, and set aside for special examination. Splitting or lumping of clusters takes place next. Splitting takes place if the conditions described below are met. Lumping occurs between the NCLST closest pairs of cluster centers that are less than  $\theta_C$  apart where NCLST is a control parameter. The process control parameters, NCLST,  $\theta_C$  and  $\theta_N$ , as well as others, are supplied by the data analyst.

After each lumping of splitting, the updated set of average points is used as the set of cluster points for the next iteration. Several statistics of the data structure are calculated and printed out.

The partitioning can be and has been done with respect to a variety of measures of similarities of data points to cluster averages. The measures of similarity used thus far are:

- (1) Normalized dot products between data points  $\{x\}$  and cluster averages  $\{m\}$ , where the normalization is with respect to the magnitudes of the means and the data points. This can be expressed as  $(\vec{x} \cdot \vec{m}) / (|\vec{x}| |\vec{m}|) = \cos(\angle \vec{x}, \vec{m})$ .
- (2) The dot product between the data point and the cluster averages. This can be written as  $\vec{x} \cdot \vec{m} = |\vec{x}| |\vec{m}| \cos(\angle \vec{x}, \vec{m})$ .
- (3) Euclidean distance squared. This can be written as  $||x - m||^2 = x \cdot x - 2x \cdot m + m \cdot m = (x - m)(x - m)'$
- (4) Mahalanobis distance, which includes Euclidean distance as a special case, which can be written as  $(x - m) W^{-1} (x - m)'$ , where  $W$  is the pooled covariance matrix, or the sum-of-products-within matrix.

As would be expected, these different measures of similarity result in different clusterings of a given set of data points. As can be seen from the various equations describing these measures of similarities, there is also considerable similarity between them. The normalized dot product measures distances only in angles between vectors. For this reason it is quite sensitive to the selection of the origin with respect to which these angles are to be measured. The dot product is not only



sensitive to the selection of origin, but it is also sensitive to the magnitude of the vector data points and cluster averages as well. The Euclidean distance does not depend on the choice of origin and can be viewed as using an additive normalization of the dot product measure of similarity to make it independent of the origin. Euclidean distance is invariant with respect to orthogonal transformations (i.e., rotations) of the data. The Mahalanobis distance is sensitive to the covariance of the data points around the various cluster centers and is invariant with respect to linear transformations of the data, as well as invariant to the position of the origin.

The division of a single cluster into two clusters in ISODATA, which we call splitting, involves first the evaluation of the desirability of dividing the cluster into two clusters, and secondly, a procedure for doing this splitting. In the original ISODATA algorithm, splitting was performed by setting an arbitrary process control parameter  $\theta_E$  and then evaluating each cluster on the basis of whether the maximum standard deviation along any of the dimensions for each of the clusters exceeded  $\theta_E$ . If  $\theta_E$  was exceeded, the cluster was split. Certain problems result when this is done. In particular, it is possible to select the value of  $\theta_E$  such that a cluster is split and then at a later time the two resulting clusters recombined because the distance between the means of the two resulting clusters was too small relative to the value of the parameter  $\theta_C$  that controls when two clusters are to be recombined. The dependence of the evaluation only on one dimension was also felt to be inadequate.

A new procedure now programmed with the ISODATA algorithm performs a trial splitting for each of the clusters. This new splitting criterion functions as follows:

- (1) Find that one dimension among the original coordinates of the data having the largest standard deviation about the mean of the cluster.
- (2) Sort the data into two subsets--a subset consisting of all patterns having a value larger than the mean in that one coordinate, and a subset consisting of all patterns having values smaller than the mean in that one coordinate. (Note that a comparison of one component of the data vector with the threshold is all that is required for this step.)
- (3) Find the means of these two subsets.
- (4) Use the magnitude of the vector difference between these two means as an approximation to the distance that would exist between the two cluster centers resulting from the split. (It is an approximation because the effect of the patterns in the other clusters is not taken into account.)
- (5) Compare this magnitude with the threshold  $(1.1)\theta_C$  and split the cluster if that threshold is exceeded. The threshold  $\theta_C$  is the parameter that determines when two clusters are to be

combined into a single cluster (lumping). The advantage of the new splitting criterion is that it is now a global splitting criterion in the sense that one measures the distance between the new cluster means after splitting using all of the dimensions rather than just evaluating the cluster on the basis of the largest standard deviation in any one dimension. It has the further advantage that it will make possible, although this has not yet been implemented, the selection of that cluster that will maximally decrease the squared error when split. This will be useful if the ISODATA algorithm is to be used to trace out the curve of MSE versus the number of clusters, as is done in the Singleton-Kautz algorithm.

The recombining or lumping of two clusters depends on measuring the Euclidean distance between all pairs of cluster averages and comparing this distance with a threshold  $\theta_c$ . In the past, all clusters having inter-pair distances greater than  $\theta_c$  have been recombined. In the future it may be desirable to combine that single pair of clusters that minimally increases the squared error. This would be simple to do because the sum-squared error is a function only of the overall mean, the two cluster means that are being considered for recombination, and the number of patterns in each cluster. If this were done, it would result in the complete elimination of the process parameters that have been used to control the ISODATA process. In certain cases it seems that removal of these parameters from consideration would be useful. In other situations, when we wish only to use the magnitude of the distance between the cluster centers to determine the number of clusters, it may be desirable to retain  $\theta_c$ .

#### Output From Computer Programs

Given that we have performed the clustering of a body of data, there remains the question of what particular fact about that clustering we wish the computer to print out for our further examination. We have found that the averages of the clusters, a list of the data points in each cluster, the distances between cluster centers and some statistics on the within-cluster spread versus the between-cluster spread are particularly helpful.

### III NEW INFORMATION RELATING TO MSE CLUSTER-SEEKING TECHNIQUES

Two recent results are:

- (1) That the curve of the MSE versus the number of clusters is not convex but that it is 'star-shaped,' which is a weakened form of convexity.
- (2) That the ISODATA algorithm will converge to a partition that is not a local minimum.

#### The Shape of the MSE Curve

Dr. Richard Singleton has been able to show\* that the curve displaying the MSE versus the number  $K$  of clusters is not convex. The counterexample he obtained is shown in Fig. 7.2. He has been able to show, however, that while the curve is not convex with respect to all possible pairs of points, it exhibits convexity with respect to those pairs of points having as one member of the pair either  $K = 1$  or  $K = N$ , where  $N$  is the number of data points. This form of weak convexity has been described previously and labeled "star-shaped." (See Bruckner and Ostrow, 1962, for a further discussion of star-shapedness.)

It is worth noting that, at least in appearance, the weakening of the convexity of this curve to star-shaped form does not appear to allow the MSE vs.  $K$  curve to be very non-convex. In the future we hope to use the star-shapedness of the MSE vs.  $K$  curve in evaluating an empirically obtained MSE vs.  $K$  curve. We would test the star-shapedness of the curve and when, for a particular value of  $K$ , the MSE ( $K$ ) violates this star-shaped condition, we would attempt to find a new partition such that the curve becomes star-shaped.

#### Convergence to Non-Local Minima

The representation of a one-dimensional data set as a contour-map-of-SSE ( $\theta_1, \theta_2$ ) allows us to investigate the dynamics of a "simple" ISODATA process (one without splitting or lumping). This plot is shown in Fig. 7.3. This representation gives the value of the sum of the squared error as a function of the position of two thresholds placed along the real line for the data shown in Fig. 7.4. In using this representation we use the knowledge that the convex hulls of an MSE partition cannot intersect. The tracks shown on the contour plot show how this "simple" ISODATA algorithm shifted thresholds from iteration to iteration. In "simple" ISODATA we used cluster averages obtained from one iteration to define the threshold positions for the next iteration, which in turn defined the cluster averages for that iteration. We see that this "settling process" does not always find even a local minimum of the sum-squared error surface but that it may (owing, we believe, primarily to the discreteness of the data) stop on a "shelf" in the SSE function fairly remote from a local minimum point of the sum squared error surface.

---

\* R. Singleton, internal SRI memorandum, June 1966.

The contour plot also illustrates the existence for this data set of two minima of the sum-squared error function for three clusters. Using the plot, we have obtained examples having all four combinations of one or two minima for two clusters and one or two minima for three clusters. At a future time we expect to use this plot to help us further in examining the relationship between the Singleton-Kautz algorithm and ISODATA.\*

---

\* The normal Singleton-Kautz algorithm run on this data and it did find the MSE partition. We will run the regular ISODATA program on this data shortly and we will give the results in the final version of this paper.

#### IV DATA SETS

In this section we describe sets of data so constructed that we believe that they will bring out the sensitivities of cluster-seeking techniques that are applied to them. These data sets can, we believe, test the power of cluster-seeking techniques to suggest structure in data. These data sets should also be useful in interpreting the results of clustering, since similar results on data of known structure might indicate a similarity in data structure between this data and data of unknown structure.

We have designed data sets to embody many conventional assumptions regarding data. In the first several sets of data it is most convenient to describe these assumptions in statistical terms.

Data Set 1 consists of a mixture of normal distributions of varying means, with each distribution having as covariance matrix the same scalar multiple of the identity matrix. (See Fig. 7.5)

Data Set 2 has the same mean values as Data Set 1, with the covariance matrices being the same for all clusters but no longer diagonal. (See Fig. 7.6)

Data Set 3 uses the mean values of Data Set 1 with different covariance matrices for each cluster. (See Fig. 7.7)

Data Sets 4 and 5 have characteristics similar to Data Set 3. Variation in a few dimensions of each cluster is low, but there is very high variation in the other dimensions. These data sets are meant to relate to data in which some measurements are very important under some conditions while other measurements are very important under other conditions. Cluster-seeking techniques ultimately should be able to isolate each underlying distribution by finding those dimensions that are of small variability. (This data can be viewed as measuring the technique's ability to cluster data points and variables simultaneously.) (See Fig. 7.8)

Data Set 6 tests for the cluster-seeking technique's ability to deal with variations in the size of clusters in different regions. (See Fig. 7.9)

Data Set 7 tests sensitivity to local variations in the data structure. In this data set the small blob of data points isolated from the main string by a region of practically zero pattern density is the important feature of the data. Minimum squared error partitions assuming identical covariance matrices for all distributions will, in general, not find the small group until the large group has been broken down into many small groups. (See Fig. 7.10)

Data Set 8 tests sensitivity for overlapping mixtures of Gaussian distributions. (See Fig. 7.11)

\* At this time only Data Sets 1,2,3,7,12,16,18,19, and 20 exist as a set of data points. The other data sets will be generated in the near future.

Data Set 9 should be sensitive to cluster-seeking techniques that look for non-linear, essentially one-dimensional data embedded in a high-dimensional space with mixed data populations. An example of a process that might generate this kind of data might be data derived from a particular word in the English language spoken many times by each of ten different speakers. If measurements are made on this word over a number of instants of time, the word itself can be viewed as a trajectory in some data space. Since there is no guarantee that even the same speaker, speaking the same word, will say it in the same way, particularly if one attempts to vary the environmental conditions under which words are spoken, it is helpful to be able to break apart words that are spoken differently and yet still be able to combine words that are spoken very similarly. (See Fig. 7.12)

Data Set 10 should be sensitive to those techniques that seek to isolate patterns into clusters, based primarily on the absence of patterns between clusters rather than on variability within clusters. (See Fig. 7.13)

Data Set 11 examines the sensitivities of techniques to particular kinds of constraints placed on the data. In this case, the constraint is that the data all lie on a spherical hypershell.

Data Set 12 consists of uniformly distributed random data. It provides a good test for the sensitivity of techniques to structure within data. If it is not easy to tell from the output of a program the difference between uniformly random data and the clustered data of Data Set 1, then we would have to assume that the particular technique being tested would probably be extremely difficult to interpret without further information being provided by the program. (See Fig. 7.14)

Data Set 13 is a collection of Gaussian distributions whose means lie in a two-dimensional space and with data points in a three-dimensional space. The data points have been rotated so that they lie in a three-dimensional subspace of a six-dimensional space. This data set provides a means for evaluating our ability to interpret results from high dimensional data when that data can be exactly characterized in a lower dimensional space.

Data Set 14 is very similar to the preceding one, but instead of a simple rotation into a higher dimensional space, a non-linear transformation was used so that linear techniques like principal components will not help much. (See Fig. 7.15)

Data Set 15 consists of data generated from complicated models plus noise, in order to see if we can recover hints as to the nature of the model. These data sets are probably closer to those obtained from a scientific experiment in which we have only a vague idea as to the underlying processes and wish to use the cluster-seeking technique to suggest what the underlying processes might be.

Data Set 16 consists of one-dimensional data mentioned earlier that is known to have certain characteristics with respect to minimum squared error (See Fig. 7.4). It has been added so that studies can be made of the dynamic process by which various cluster-seeking techniques arrive at a particular partition of the data.

Data Set 17 consists of five-dimensional data for which an attempt has been made to minimize the information obtained from a marginal distribution along any dimension or pair of dimensions in a scatter plot and so situated that a principal components analysis gives little information. The data itself is well-clustered in the sense that for each cluster, within cluster deviations are very small with respect to the distance between a cluster and its closest neighbor.

Data Set 18 is the historic Fisher-Kendall data set describing four measurements made on three species of Iris. This data is included to aid the comparisons between techniques that have been developed over a considerable period of time, since this data set has been used by a number of experimenters. It is, however, a fairly simple set of data.

Data Set 19 is a large body of data consisting of 20 measurements on 1000 data points, provided by Dr. Bernard Glueck of the Institute for Living. This data does not have well-known structure and is probably rather complicated. It is a test not only of our ability to interpret the data, but it also provides a good evaluation of the technique's capabilities with respect to large data sets of real data with relatively high numbers of dimensions.

Data Set 20 consists of 122 measurements made on 97 species of bees, by Michener and Sokal, and has been included to provide a data set in which the number of measurements exceeds the number of dimensions.

It is hoped that these data sets will provide a sufficient experimental exercising of a proposed cluster-seeking technique to provide a reasonably good understanding of the capabilities of this technique.

Due to the large number of data sets we discuss only Data Sets 1, 2, and 3 in comparing the Singleton-Kautz algorithm and ISODATA.

## V COMPARISON OF TECHNIQUES

The comparison of the two techniques is divided into a section dealing with verbal and graphical comparisons, a second stating analytical differences and similarities, and a third dealing with experimental results.

Assumptions. The Singleton-Kautz Algorithm and ISODATA assume that a disjoint partition of the data set with relatively homogeneous data points being placed in the same partition is useful. Homogeneity is measured by a "distance" to a cluster average. They assume that the particular distance measure that they use is valid. Particular variations of these techniques are obtained in the case of the Singleton-Kautz algorithm by modifying the criterion against which improvement in the partitioning is measured, and in the ISODATA technique by modifying the measure of similarity, and by modifying the procedure by which clusters are split and lumped. Global or local evaluating criteria can be used with ISODATA to further constrain the solution obtained. No explicit distributional assumptions are made in either of these techniques. However, it is assumed that the distance measure or the criterion used is adequate to reflect the structure of the data accurately.

Economies of Description. These cluster-seeking techniques describe those situations most economically in which isolated clusters of data exist with dimensional variation that is high in the sense that the covariance matrix of the means of these clusters is of rank nearly equal to that of the data space. These techniques are not particularly efficient in describing relatively uniform random variability that occurs within a low-order linear subspace of the original data space. For these situations the factor-analytic techniques that look at these linear subspaces seem more appropriate. They can, however, still be used in these situations to provide empirical data categories. The cluster-seeking techniques try to group patterns so that the average squared distance from cluster means is not significant. Factor-analytic techniques seek to place the data in a lower dimensional space and then retain the full variability of the data within that lower dimensional space.

Limitations. These cluster-seeking techniques, when using either a criterion or a measure of similarity corresponding to Euclidean distance, are sensitive to changes in scaling, although they are not sensitive to rotations of the data or the position of the origin. Changes in the data set that affect normalizations based on the data sets, such as the standard deviation about a mean, may modify the clustering obtained. When the ISODATA technique is used with the Mahalanobis distance it is relatively insensitive to the scaling of the data. The results of using these techniques are frequently difficult to interpret because these results have a large number of degrees of freedom. Therefore, any simple interpretation usually could arise from a great variety of data sets. Hence these techniques, when used to obtain too a simple description may provide little interpretive discrimination between data sets.



(This same criticism holds for most techniques based solely on the covariance matrix). Complex descriptions that more accurately reflect details of the data are apt to be confusing.

Invariances. The Singleton-Kautz algorithm is invariant with respect to orthogonal transformations and translations of the data. The ISODATA technique using Euclidean distance is also invariant with respect to orthogonal rotations of the data and translations of the origin of the data. If the Mahalanobis distance is used, then ISODATA is largely invariant with respect to linear transformations as well as with respect to translations of the data. These techniques both tend to produce different results if individual data points are deleted from the data set, particularly if these data points are "outliers" or "wildshots."

It is well to reiterate that different kinds of data may be invariant with respect to the clustering procedure in that the clustering procedure may not be sensitive to the ways that these data sets vary. If the particular variability is important, then a technique has to be developed that is sensitive to this variation. For example, if scale is important and there is some natural way of defining the scale, or where there is a desire to weight certain variables more heavily, then invariance with respect to scale would not be a desirable feature for a technique

Extensions to Different Problems. ISODATA appears to be more directly extendable to the clustering of points around line segments or planar sections. At this time an algorithm is being programmed\* to cluster points around line segments. This algorithm uses the following notions that exist in the ISODATA algorithm.

- (1) The creation of new cluster centers (the cluster centers are now line segments).
- (2) The evaluation of the usefulness of a given line segment.
- (3) The iterative shifting of the line segment to place it in "better" position.
- (4) The combination of those line segments that can be combined without greatly reducing the information we have regarding the structure of the data.

When we have completed programming this algorithm we will investigate the desirability and the feasibility of clustering data around triangular planar sections. This would enable us to approximate mixtures of non-linear two parameter surfaces that are embedded in a hyperspace.

\* James Eusebio of SRI has done all of the programming and much of the work constructing this algorithm.

Goals. These cluster-seeking techniques have as their goal the determination of structure in data. They are sensitive to density variations in the data in the original high dimensional space. Both techniques can be used on any kind of data (including nominal data that is correctly encoded). Interpretations must correspond to the assumptions given above that were made in developing the techniques and whether these assumptions are being met by the data.

Analytical Comparison. These techniques minimize a criterion subject to certain constraints. The Singleton-Kautz algorithm explicitly evaluates and minimizes SSE as a global criterion. It is constrained in this minimization to make reassignments of single data points. It performs this minimization by hill climbing (or really, valley descending) from a variety of starting positions and then selecting for one cluster up to KMAX clusters the lowest value of SSE found in the various tries as the overall minimum.

The ISODATA technique tends toward implicit minimization of SSE by requiring that a stable partition meet the necessary condition given above. Its settling procedure is not as powerful as the single move minimization, as can be seen in Fig. 7.16. For the data of this figure either threshold satisfies the conditions for a stable ISODATA partition. Only the optimum threshold  $\theta_1$  satisfies the stopping criterion of the single move algorithm. The ISODATA technique is constrained to find that minimum squared-error partition that keeps the minimum distance between the means of all pairs of clusters greater than  $\theta_c$ .

The computation time for the "inner loop" of the Singleton-Kautz Algorithm and for the "inner loop" of the ISODATA program using Euclidean distance as its measure of similarity is approximately equal. The question of convergence per iteration remains to be examined as does the effect of large numbers of data samples and of high dimensionality of the data.

Experimentally on 225 two-dimensional data points we have observed that the Singleton-Kautz algorithm finds a partition of the data that has a SSE that is about 10 per cent lower than that of the partition found by ISODATA. We do have instances, however, when ISODATA has found a lower SSE for this same type of data. Perhaps more importantly, for many applications, we have recently noticed that the Singleton-Kautz algorithm required 15 iterations through 1000 six-dimensional data samples before it found a single move minimum for SSE. We plan to examine this question further by making a detailed comparison of the iteration by iteration reduction by these two algorithms of SSE on a variety of data sets. This has not yet been done as it requires some modification of the Singleton-Kautz program to allow the two programs to start from the same partition of the data. The evidence thus far is that the Singleton-Kautz algorithm generally finds a lower value of SSE than does ISODATA.

If a more complicated distance function is used, such as a sum of the Mahalanobis type distances, then the necessity for the Singleton-Kautz

algorithm to invert a matrix after each data sample increases the computation of its "inner loop."

Since ISODATA does not change  $W^{-1}$  until all of the patterns have been resorted, ISODATA should have lower running times with the Mahalanobis type of measure of similarity.

Experimental Comparisons. The experimental comparisons described in this paper were confined to Data Sets 1, 2, and 3. The results are summarized in Tables I, II and III and Figs. 7.17 and 7.18.

For Data Set 1 the Singleton-Kautz Algorithm and ISODATA produced identical clusterings pattern for pattern.

TABLE I - EXPERIMENTAL RUNS ON DATA SET 1

Singleton-Kautz		ISODATA	Data Means
1.9	7.9		2,8
7.9	6.5		7,7 and 8,6
9.1	1.1		9,1
2.9	2.1		3,2
6.0	3.0	← Identical to Singleton-Kautz	6,3
1.0	4.3		1,4
4.1	8.9		4,9
5.1	5.1		5,5
SSE 139.69		139.69	

For Data Set 2 the clusterings were quite similar, with only two out of ten cluster centers being very different.

TABLE II - EXPERIMENTAL RUNS ON DATA SET 2

Singleton-Kautz		ISODATA		Data Means
4.8	9.4	5.0	9.5	4,9
8.4	6.9	9.0, 6.7 and 7.5, 7.3		8,6; 7,7
9.2	1.3	9.1	1.2	9,1
4.9, 2.2 and 6.7, 3.4		5.4,	3.2	6,3
2.1,	1.7	2.5,	1.7	3,2
1.7,	7.7	2.2,	8.2	2,8
3.3,	4.7	2.0,	5.4	
-.1,	3.8	-.4,	3.4	
6.4,	5.7	6.5,	5.5	
SSE 296.27		333.26		

For Data Set 3 two cluster centers are markedly different in those regions having few data points.

TABLE III - EXPERIMENTAL RUNS ON DATA SET 3

Singleton-Kautz		ISODATA		Data Means
1.6	8.4	1.6	8.4	2,8
5.2	8.3	5.1	8.3	
5.1	3.5	5.0	3.7	
1.4	3.8	.75	3.5	1,4
11.7	.7	11.9	.5	
8.2	5.9	8.1	6.0	8,6
7.2	2.0	7.5	2.2	
2.3	5.2	3.4	5.4	
2.1	.1	4.5	.85	
-.8	2.1			
		1.1	-.2	
SSE 380.55		411.19		

The Singleton-Kautz algorithm (SKA) found a lower-valued SSE partition for Data Sets 2 and 3 than ISODATA. The positions of the clusters were almost the same in most instances. ISODATA quickly got a reasonably good partition for these data sets but was very slow in improving it. SKA found a reasonable partition almost as rapidly for these data sets and improved it rapidly. SKA is considerably easier to run, since it systematically provides values for minimum SSE for all numbers of clusters up to KMAX. However, in runs on other data that had a considerable number of wildshots and was of higher dimension, ISODATA proved to be easier to interpret and run since it was not as affected by the wildshots. That is, ISODATA rapidly increased the number of clusters until the wildshots were isolated. In this latter application, ISODATA was more effective.

The statistics of the data that the two programs provide as output are almost identical.

## VI INTERPRETING EXPERIMENTAL RESULTS

In this section we discuss

- (1) An analytical technique for examining the amount of structure in data.
- (2) The ways that graphs can usefully aid in interpreting experimental results.
- (3) An interactive computer system for analyzing multivariate data.

### Random Reshuffling of the Components of Data Points as a Non-Parametric Test of Structure in Data

Dr. James MacQueen (1966) has suggested that the random rearrangement of the values associated with each component of the set of data vectors is a way in which a non-parametric test of the amount of structure in the original data can be made. More precisely, consider an ordered set of data points in which each data point is a row in a data matrix and each variable has its values in a column in the data matrix. First, the data is clustered using the original data points and some measure--for example, SSE-- is made of the reduction in variability around local cluster means resulting from the clustering. Next, each column of the data matrix is independently, randomly rearranged. This causes the values of each variable for the data points to be randomly associated with the values of other variables from other data points, which tests if the specific associations found in the data are important. The effect of this is to create a more or less uniform distribution of data points within the rectangular hyper-parallelepiped that contains all of the data points. If the disorganization (measured by SSE) increases perceptibly on repeated trials of clustering of the reshuffled data, then it can be said that statistically the original data was more structured than would be expected on the basis of chance. In other words, if the value of SSE for the original data is at the extreme lower end of all of the sample values of MSE obtained by this random reorganization of the data, one could say, with some statistical confidence, that the original data was structured.

This seems an exceptionally important concept in the evaluation of the results of cluster-seeking techniques. Its primary disadvantage lies in the recomputation required, since first you must randomize the data points and second you recluster all of the randomized data. As the cost of computer analysis is further reduced, this disadvantage is reduced, particularly since an interpretation that the data has structure may be greatly strengthened by this test.

### Graphical Presentations

In using graphs to aid in the analysis of data, several points seem particularly important. These graphs should include presentation

of residuals and means of limiting the subset of variables and/or data points plotted, or allow use of special symbols. Five such graphs are described\* in the following:

- (1) In any projection of a set of data points down onto a lower dimensional space it seems important to know the amount of the residual variation remaining that is not shown in the plot itself. Such a plot is shown in Fig. 7.19, in which the data is plotted with respect to a line in the hyperspace. Points are plotted with respect to two coordinates, one their distance along the line in terms of the perpendicular projection down onto the line, and, secondly, the distance they lie from the line measured perpendicularly.
- (2) We can project high dimensional data down onto an arbitrary plane. The distance perpendicular to that plane, i.e., the residual variation, is indicated in the plot by the size of the symbol. (See Fig. 7.20.)
- (3) The "metroglyph" suggested by Edgar Anderson (1957) shows either a small, medium, or large amount of residual for three additional variables. This symbolology can be grasped quite quickly by eye. A metroglyph is shown in Fig. 7.21.
- (4) We can project data points down onto an arbitrary plane without indicating the residuals. If this is done, however, it seems important to position this plane or the line meaningfully and to restrict those points that are plotted on this graph to those having small residual variation. (See Fig. 7.22.)
- (5) The graph of  $MSE(K)$  vs.  $K$  is extremely useful for MSE clustering algorithms. In Fig. 7.23 we show the difference in this curve between Data Set 1, clustered data, and Data Set 12, uniformly random data.

#### An Interactive Computer System for Analyzing Multivariate Data

In a project presently underway at Stanford Research Institute\*\* we are making use of an interactive computer to give us considerable convenience in selecting and modifying the point of view from which we

---

\* For a discussion of a wider variety of multivariate plots see Ball (1967).

\*\* David Hall of SRI and the writer have done the planning on this project together. Mr. Hall and Dan Wolf have done the computer system design and the programming. The project is supported by Air Force Contract AF 30(602)-4196 under the technical cognizance of the Information Processing Branch of Rome Air Development Center.

examine our data. Parts of the computer system are shown in Fig. 7.24. We will have available data manipulation programs that would allow us to modify the scaling or the variables actually used in describing a data set, and statistical routines including principal components and the like. Eventually, we will have available statistical routines for testing hypotheses. We have cluster-seeking techniques for finding good place within the data to look, and a section in which it would be possible to create data artificially in order to test a particular model or to generate data from a model with which experimental data can be compared. Perhaps most importantly we will have a large variety of graphical presentations that will allow a person to explore the data points as nearly as possible in their proper perspective in the hyperspace in which they lie. It is our intention that we will be able to do this with considerable convenience. If this occurs, we expect to be able to far surpass what the human being is able to do with a series of two-dimensional plots, since we will be able to guide the computer into those positions that will give us the most "information."

## VII CONCLUSIONS

For systematic analysis of relatively clean data, where the finding of the MSE partition for small numbers of clusters is a reasonable goal, the Singleton-Kautz algorithm appears to find partitions that have lower values of SSE than ISODATA. From past experience with other data, ISODATA appears to be superior for noisy data, where the goal is quick isolation of the principal modes of the data with exclusion of outliers.

The program implementing the Singleton-Kautz algorithm is easier to use in a batch-processing computer. We feel that ISODATA may prove easier to use in an interactive computer in which the judgment of the operator is used in lumping, splitting and evaluating clusters.

The relative speed of convergence of the two algorithms to an MSE partition apparently depends to a greater degree than we had expected on number of patterns and number of dimensions. This aspect of the comparison must await further experimental investigation.

For finding partitions that minimize the sum of Mahalanobis type distances it appears at this time that ISODATA would be computationally more rapid.

Interpretation of the results of these clusterings is by no means easy. Several different ways of presenting the data are described and an interactive display-oriented computer system for analyzing multi-variate data is discussed.

At this time we see the two most important goals of cluster-seeking techniques as being:

- (1) To describe the data as simply as possible, consonant with the user's need for accuracy.
- (2) The evaluation of the degree to which structure exists in a body of data.



## VIII ACKNOWLEDGMENTS

This paper describes techniques and ideas contributed by many individuals. I would particularly like to thank David Hall, with whom I have worked over the last three years, both on ISODATA and on the interactive computer system. I would also like to acknowledge the contributions of Dr. Charles Dawson, James Eusebio, and Dr. Richard Singleton, my colleagues at Stanford Research Institute. In addition, I would like to thank our statistical consultant, Professor Ingram Olkin of Stanford University's Statistics Department, and Prof. Thomas Cover of Stanford University's Electrical Engineering Department. Conversations from time to time with Herman Friedman, Dr. Edward Forgy, Dr. James MacQueen, and Jerrold Rubin have added considerably to my understanding of cluster-seeking techniques.

The work described in this paper has been primarily sponsored by the Information Sciences Branch of the Office of Naval Research under Contract Nonr 4918(OO) and by internal funding by Stanford Research Institute. The work on the interactive computer system for multivariate data analysis has been sponsored by the Information Processing Branch of Rome Air Development Center under Contract AF 30(602)-4196.

## REFERENCES

Edgar Anderson, "A Semigraphical Method for the Analysis of Complex Problems," Vol. 2, No. 3, Technometrics, August 1960, pp. 387-391, as it appeared in the Proceedings of the National Academy of Sciences, Vol. 13, pp. 923-27 (1957).

Geoffrey H. Ball, "Data Analysis in the Social Sciences--What About The Details?" Proc. of 1965 Fall Joint Computer Conference, Vol. 27, Part I, (1965).

Geoffrey H. Ball, "A Set of Multivariate Graphs Applied to the Analysis of Real Psychological Data," Final Report on Institute Sponsored Research Project 186531-157, Stanford Research Institute, Menlo Park, Calif., (January 1967).

Geoffrey H. Ball and David J. Hall, "ISODATA, A Novel Method of Data Analysis and Pattern Classification," Stanford Research Institute Technical Report, (1965).

Geoffrey H. Ball and David J. Hall, "ISODATA, An Iterative Method of Multivariate Data Analysis and Pattern Classification," IEEE International Communications Conference, (1966).

A. M. Bruckner and E. Ostrow, "Some Function Classes Related to the Class of Convex Functions," Pac. J. of Math., Vol. 12, No. 4, (1962).

Edward W. Forgy, "Improving Classification Systems for Multivariate Observations," University of California, Los Angeles, (1966).

H. P. Friedman and J. Rubin, "On Some Invariant Criteria for Grouping Data," Biometrics Conference, Upton, L.I., New York, (28 April 1966).

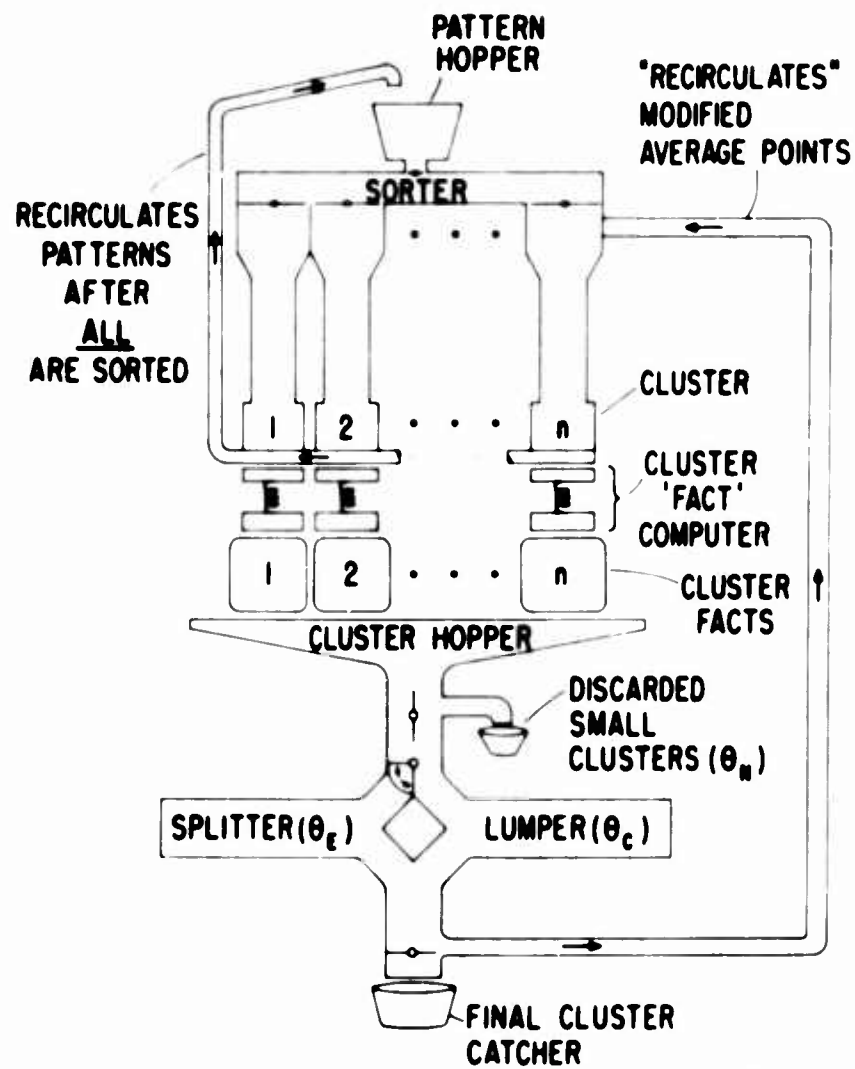
J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Working Paper No. 96, Western Management Science Institute, University of California, Los Angeles (March 1966).

J. MacQueen, "On the Asymptotic Behavior of K-Means," University of California, Los Angeles, (1966).

George S. Sebestyen and Jay Edie, "Pattern Recognition Research--Final Report," Litton Systems, (1964), AD 608-692.

George S. Sebestyen, "Automatic Off-Line Multivariate Data Analysis," Proceedings of the Fall Joint Computer Conference, 1966, Spartan Books, (November 1966).

Lawrence Stark, Mitsuharu Okajima, and Gerald H. Whipple, "Computer Pattern Recognition Techniques: Electrocardiographic Diagnosis," Communications of the Association for Computing Machinery, Vol. 5, No. 10, (October 1962).



TA-658582-1

FIG. 7.1 A PICTORIAL DESCRIPTION OF ISODATA



7.26

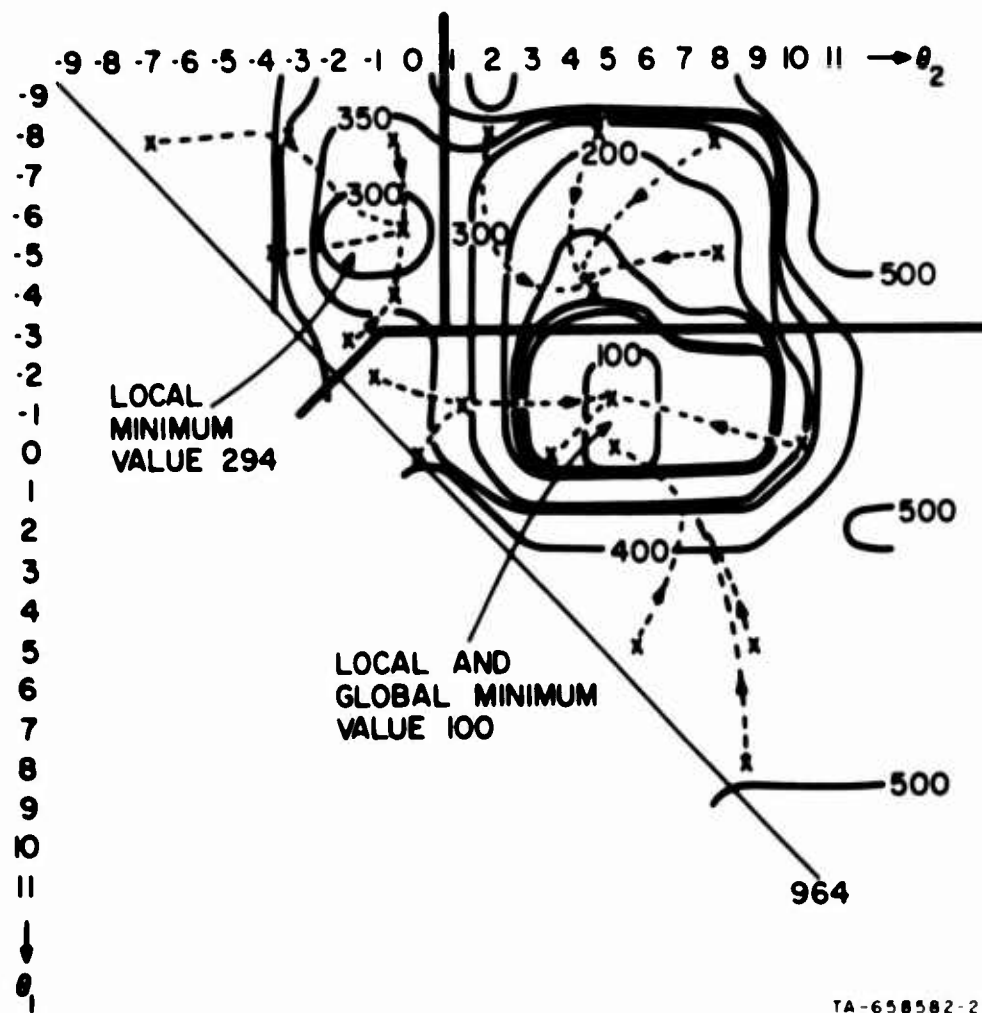
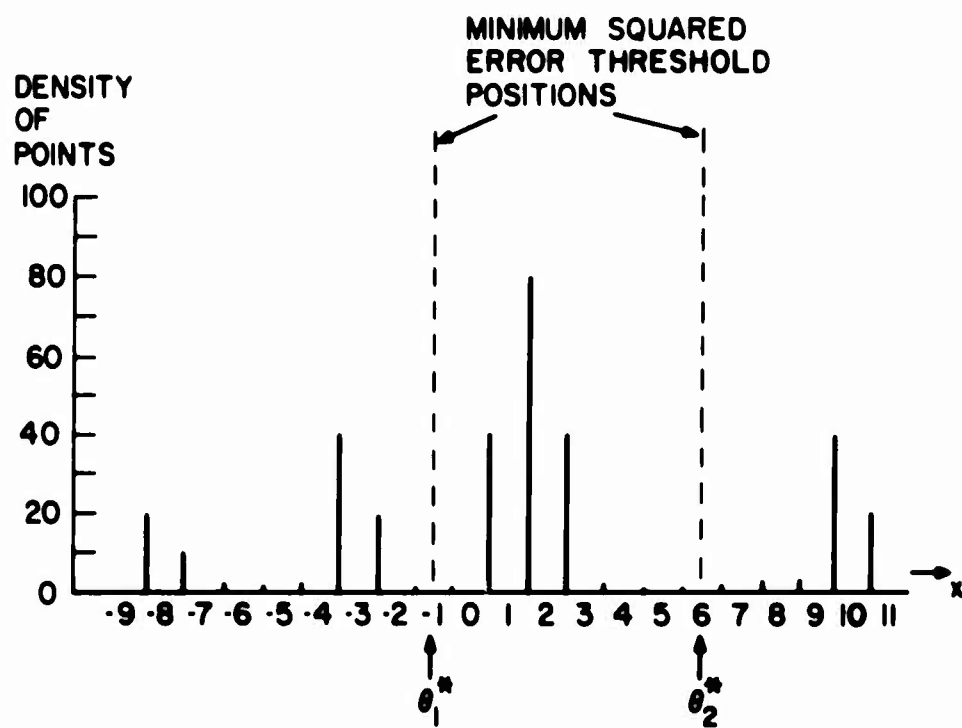
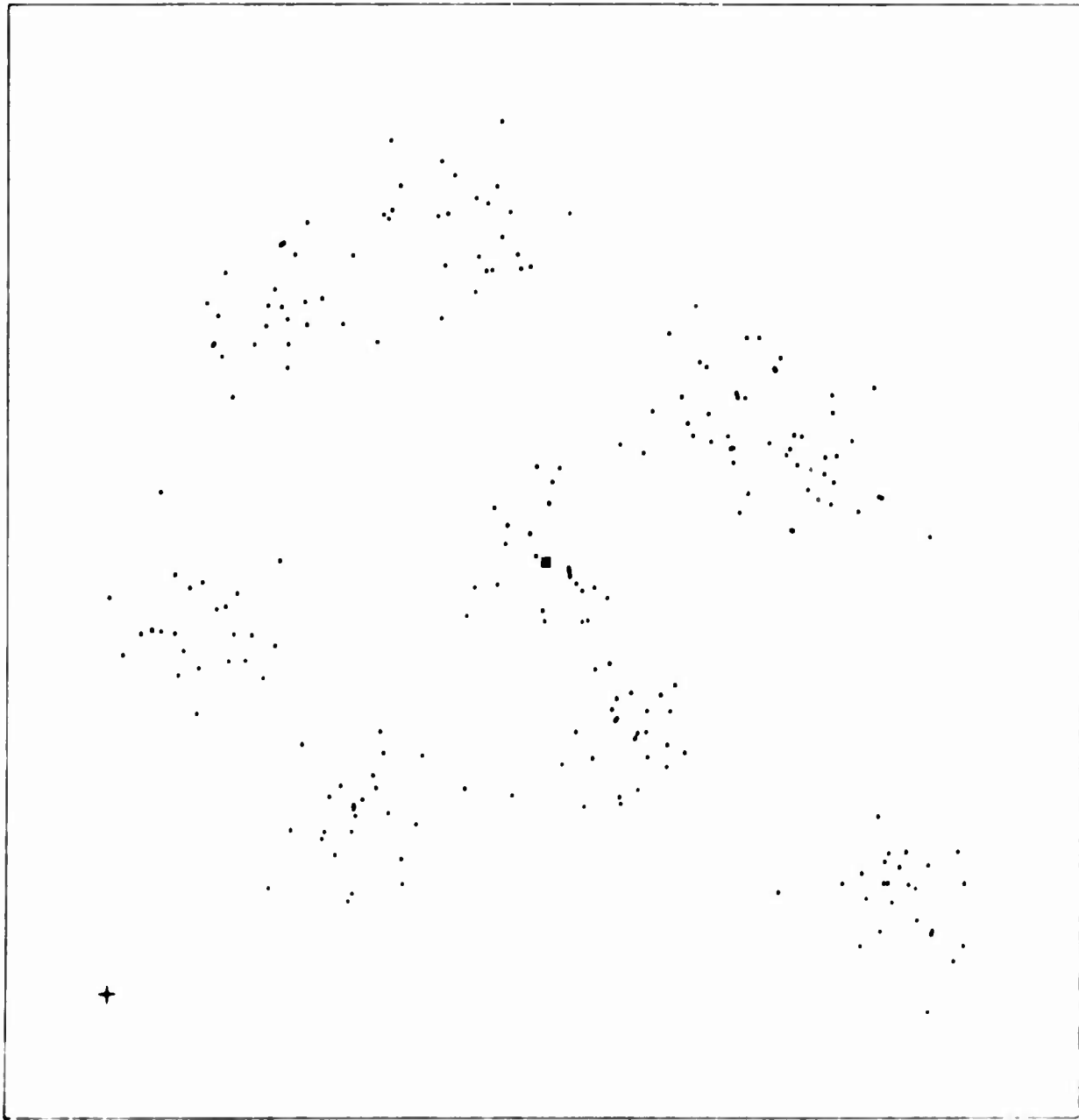


FIG. 7.3 CONTOUR PLOT OF VALUES OF SSE FOR TWO THRESHOLDS, " $\theta_1$ " AND " $\theta_2$ "



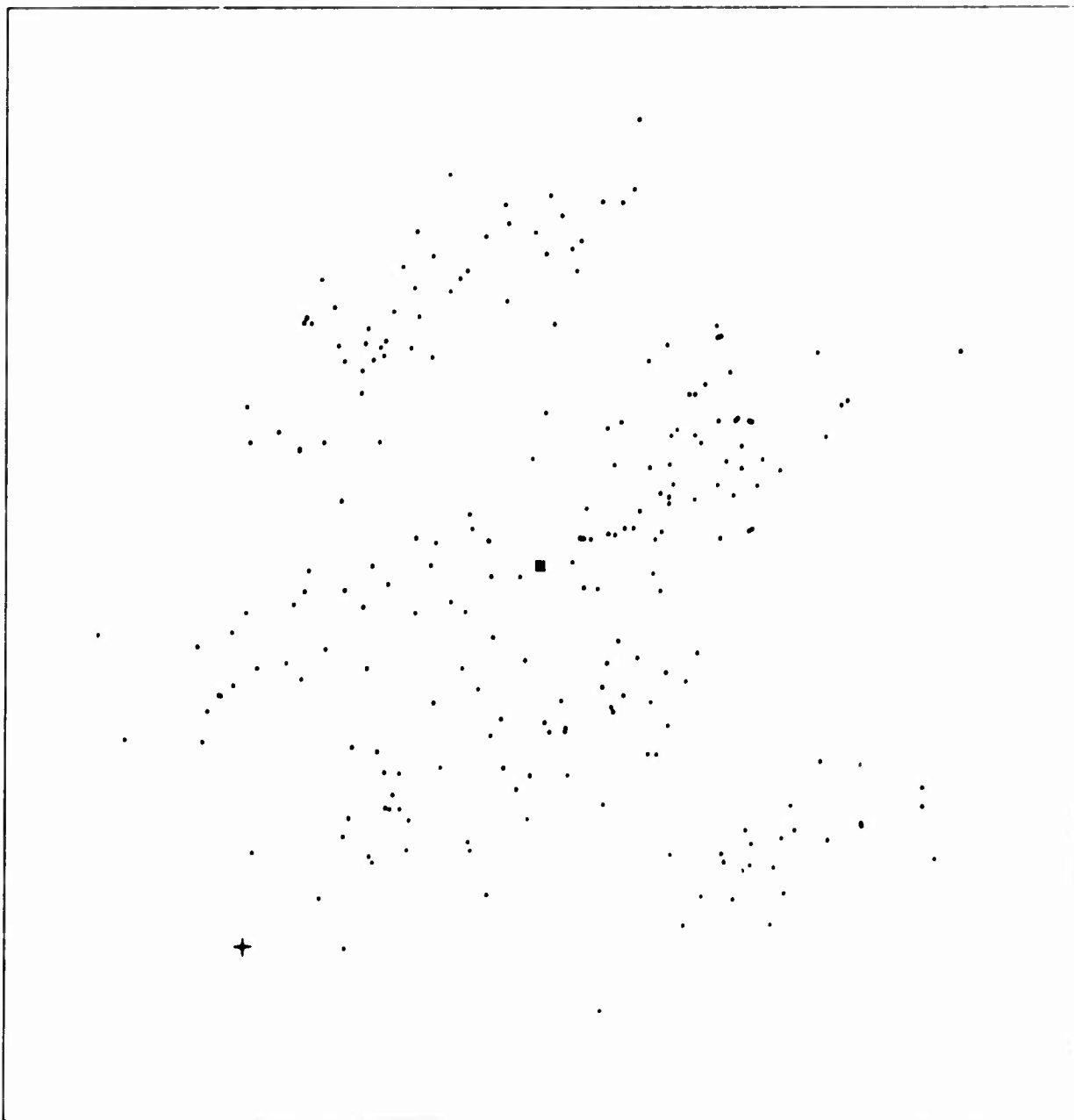
TA-658582-4

FIG. 7.4 ONE-DIMENSIONAL PATTERNS



TA-658582-5

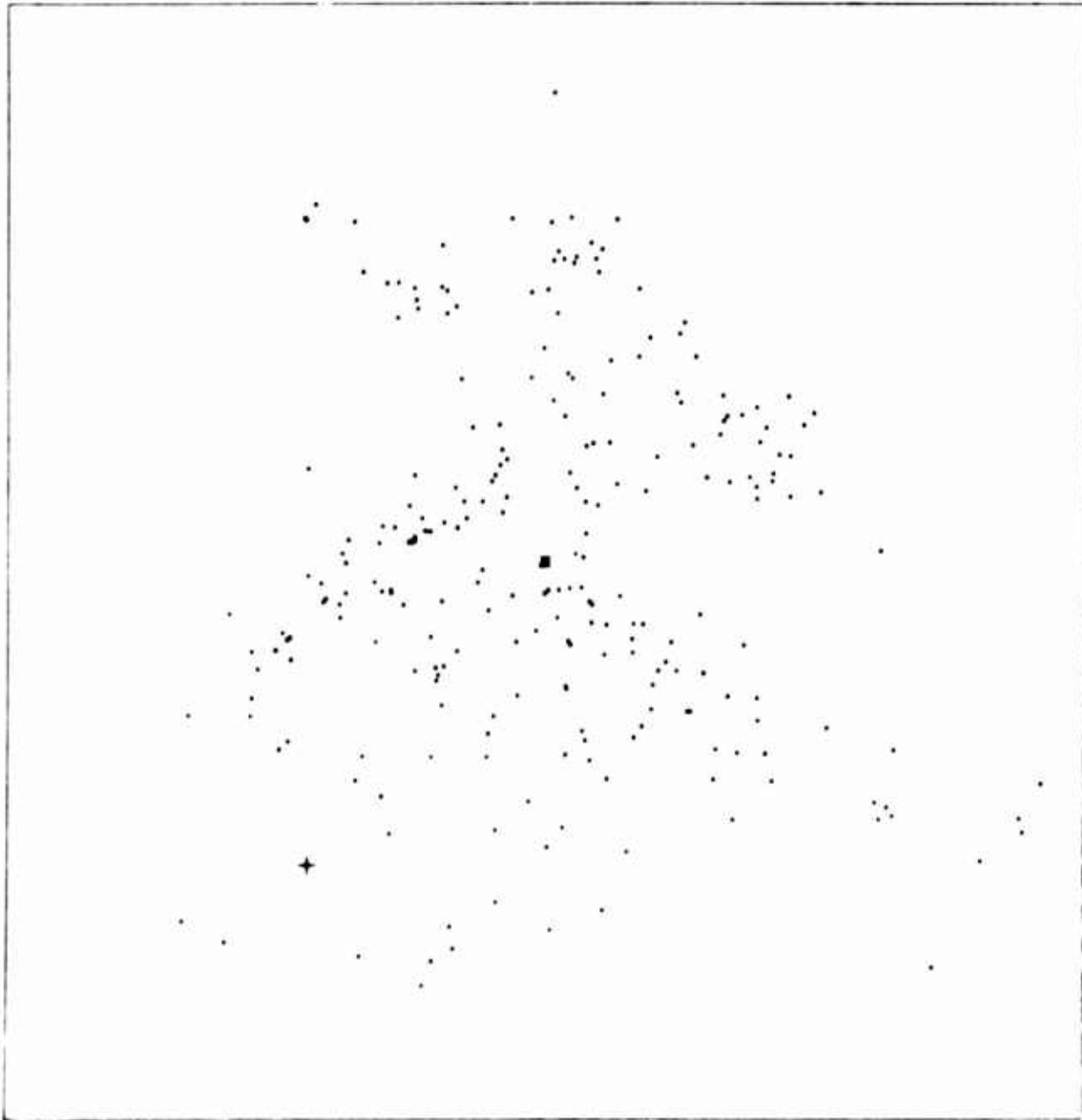
FIG. 7.5 DATA SET 1



TA-650502-6

FIG. 7.6 DATA SET 2





TA-690502-7

FIG. 7.7 DATA SET 3

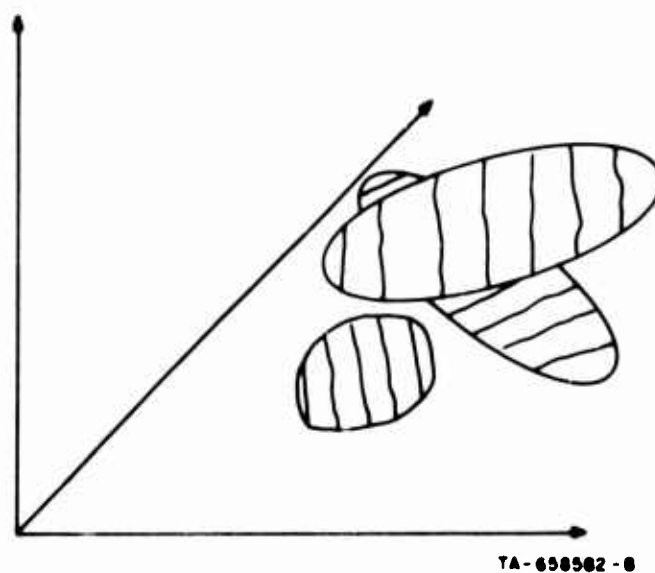


FIG. 7.8 DATA SET 4. Data set 5 is similar but in ten dimensions.

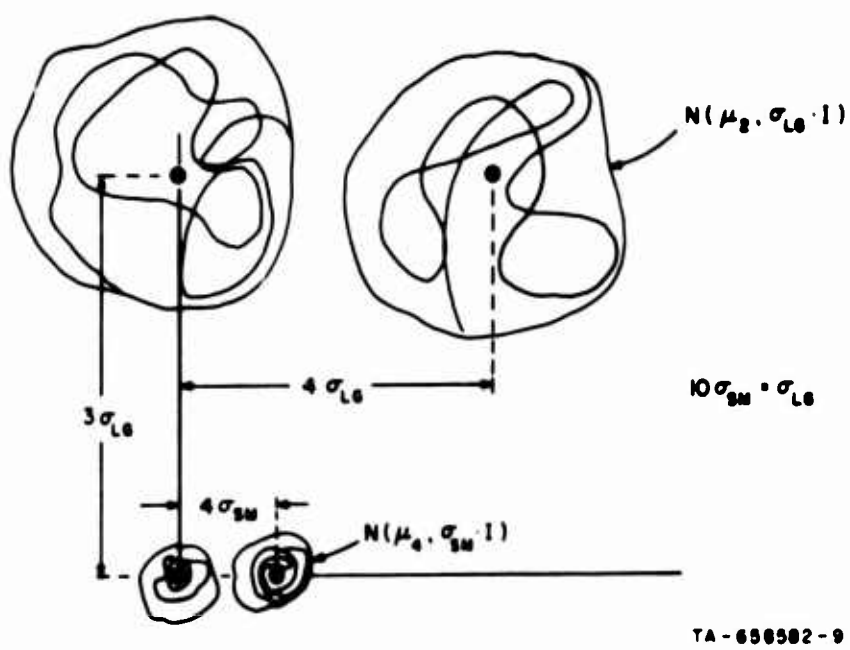
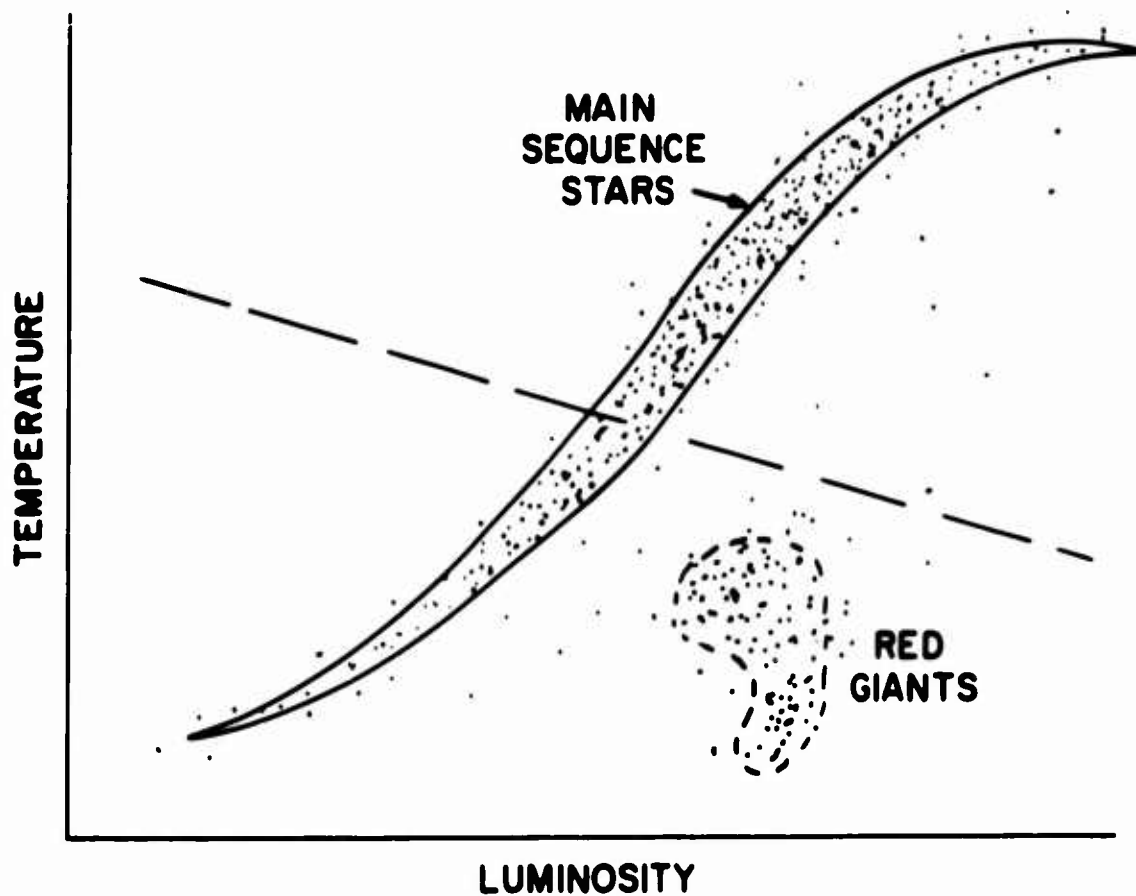


FIG. 7.9 DATA SET 6



**HERTZSPRUNG & RUSSELL DIAGRAM  
OF STARS**

TA-658582-10

FIG. 7.10 DATA SET 7 (Following Forgy)

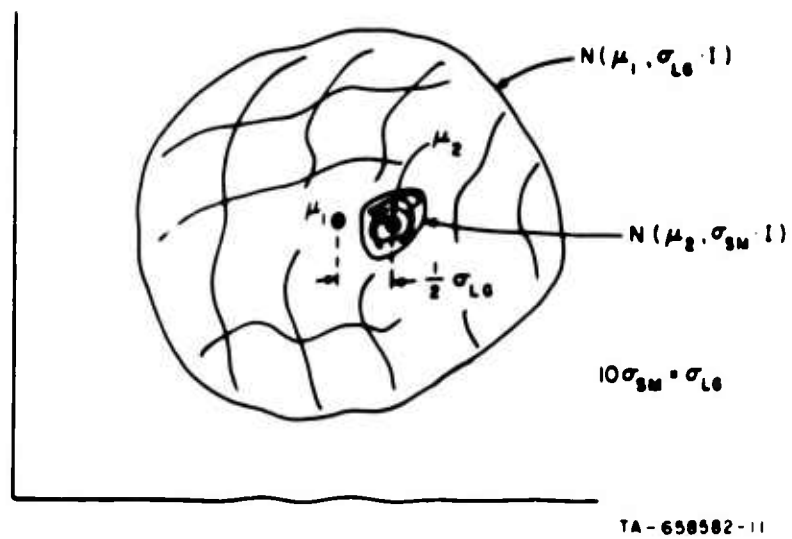


FIG. 7.11 DATA SET 8

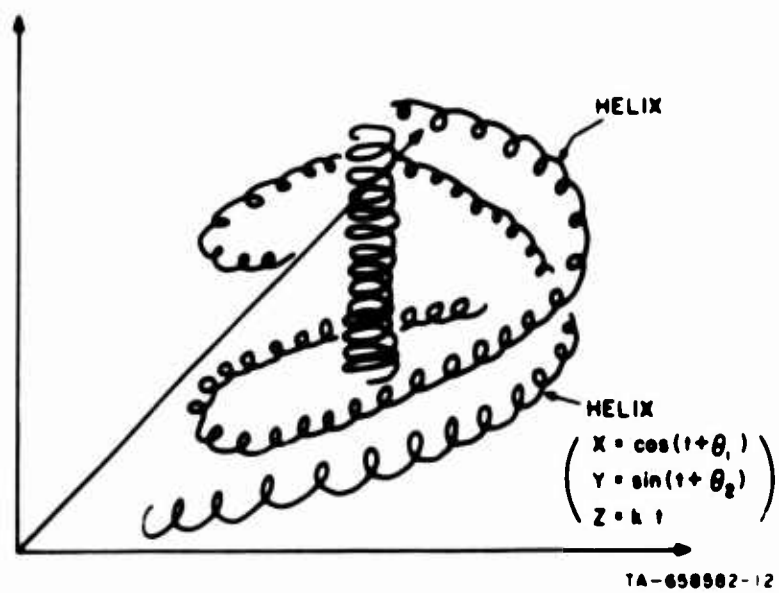


FIG. 7.12 DATA SET 9

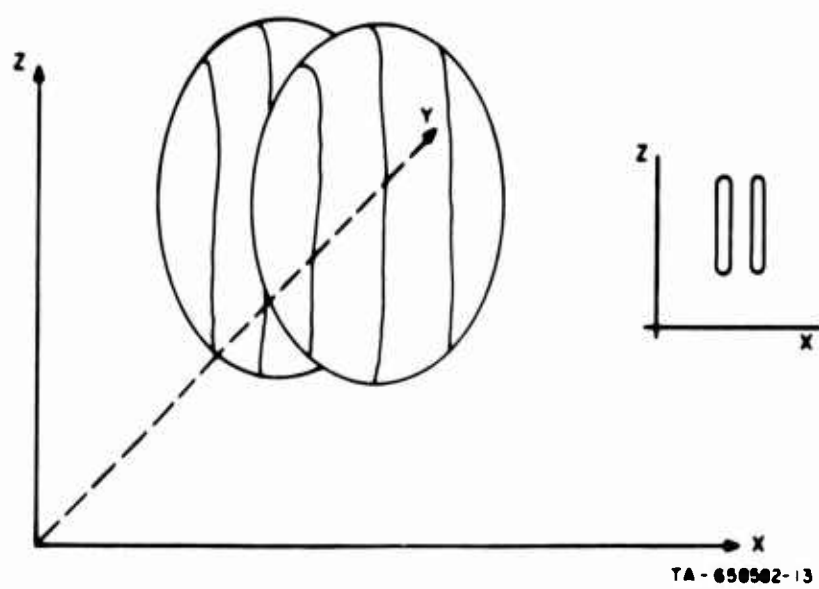
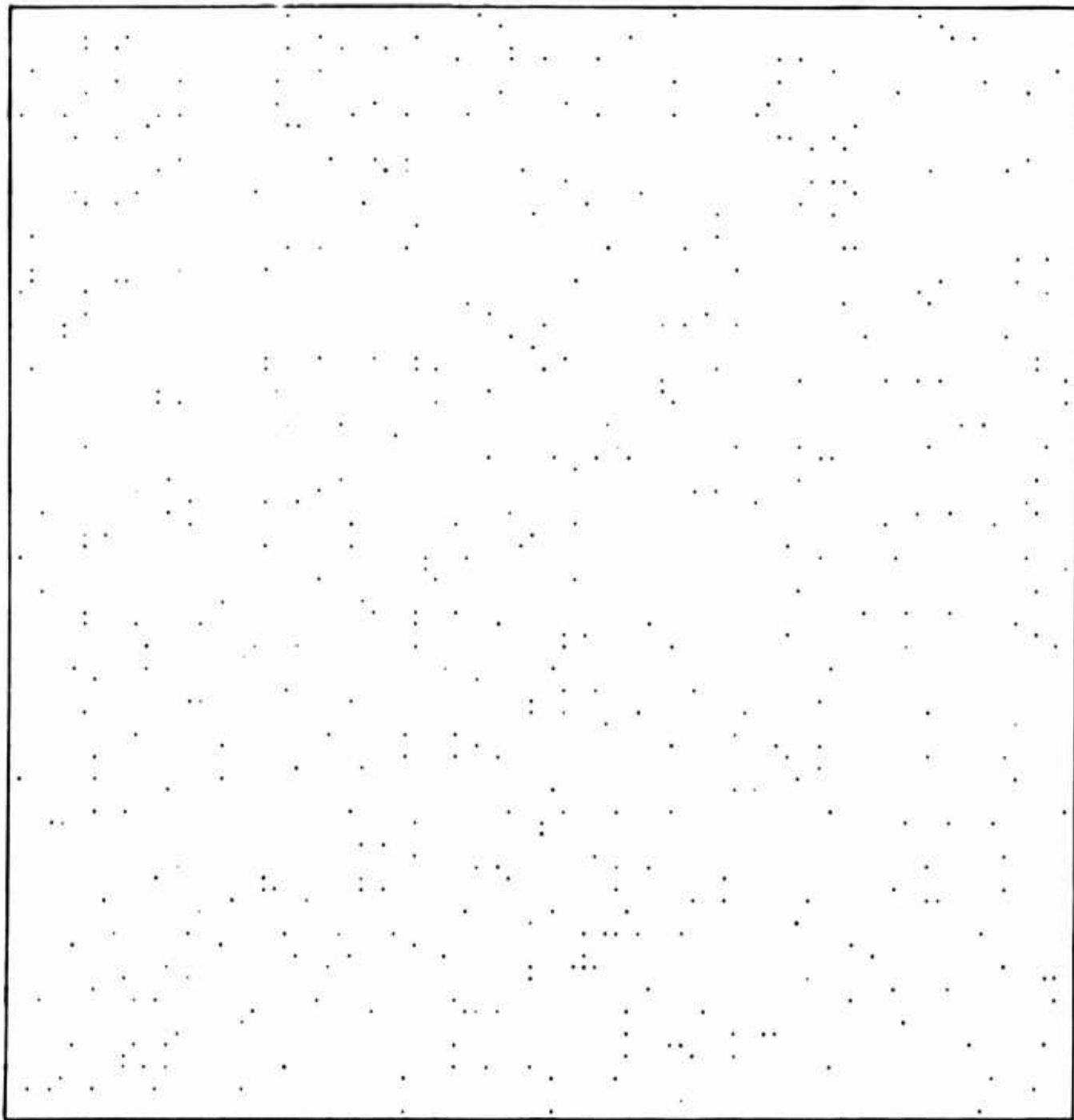


FIG. 7.13 DATA SET 10



TA-658582-14

FIG. 7.14 DATA SET 12

$$\begin{aligned}
 Y_1 &= X_1 \cdot X_2^2 / 100 \\
 Y_2 &= X_2 \cdot X_1^2 / 100 \\
 Y_3 &= \text{Noise } (N(0,1)) + X_1 \cdot X_2 \\
 Y_4 &= (X_1^3 + X_2^3) / 200 \\
 Y_5 &= (X_1^3 + X_3^3) / 200 \\
 Y_6 &= (X_2^3 + X_3^3) / 200
 \end{aligned}$$

where  $(X_1, X_2) \in \text{DATA SET 1}$ , and  $X_3$  is random Gaussian noise,  $N(0,1)$ .

TA-568582-15

FIG. 7.15 EQUATIONS FOR DATA SET 14

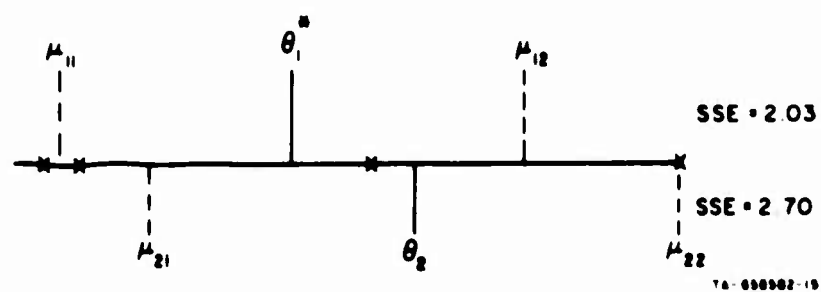


FIG. 7.16 SINGLE MOVE vs. SETTLING



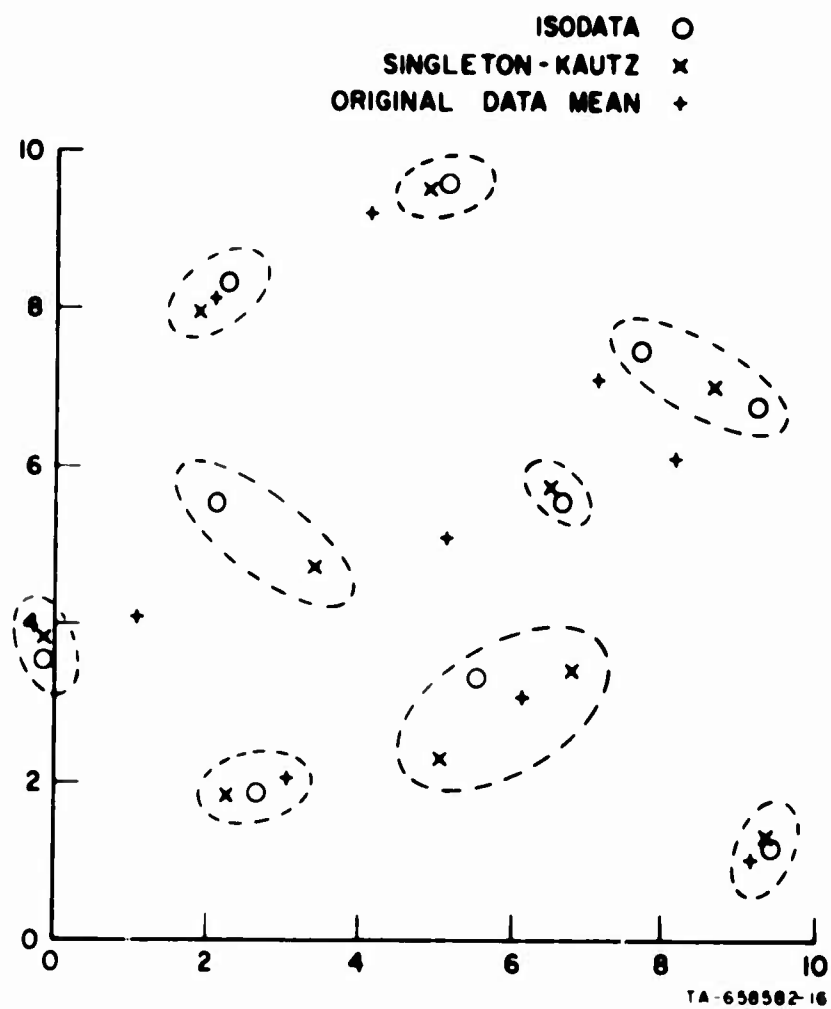


FIG. 7.17 DATA SET 2 CLUSTER MEANS FOUND

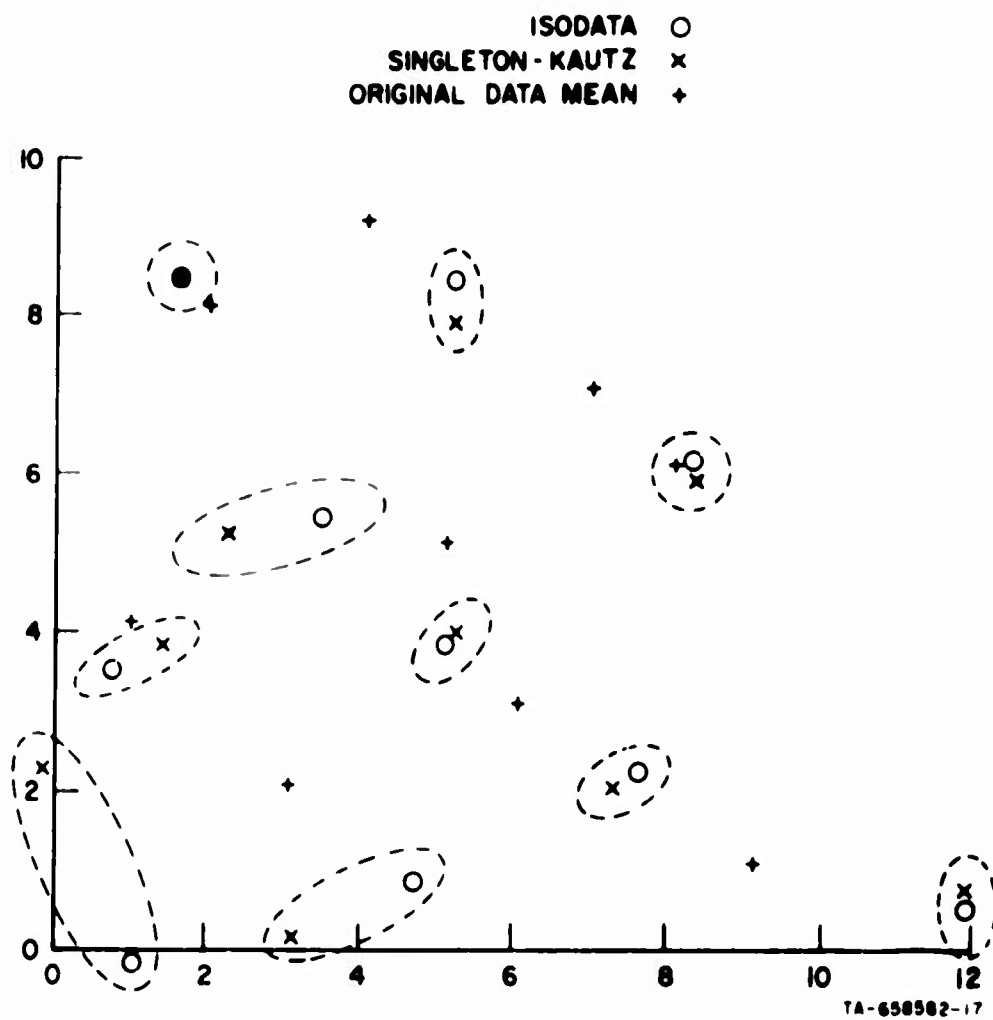


FIG. 7.18 DATA SET 3 CLUSTER MEANS FOUND

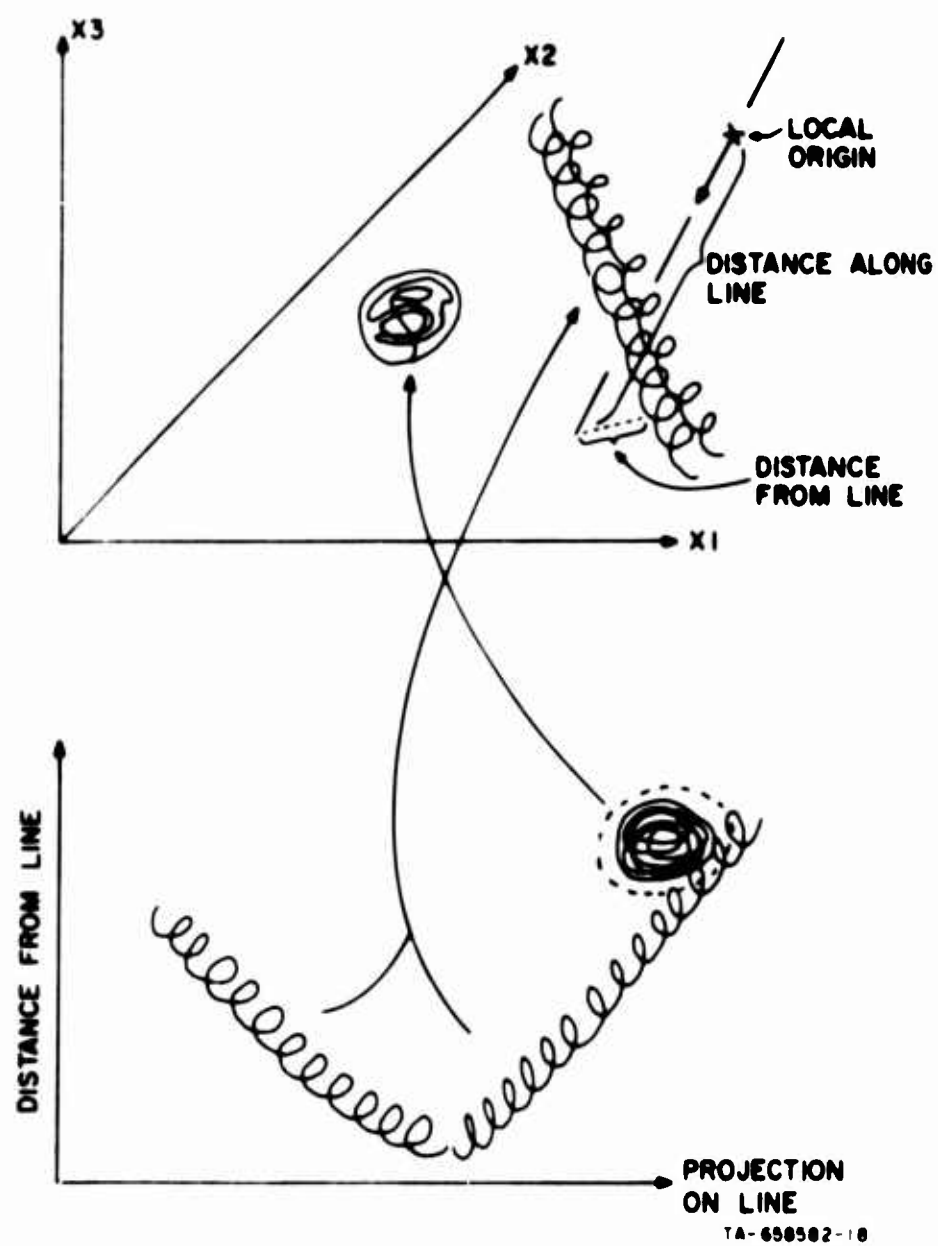
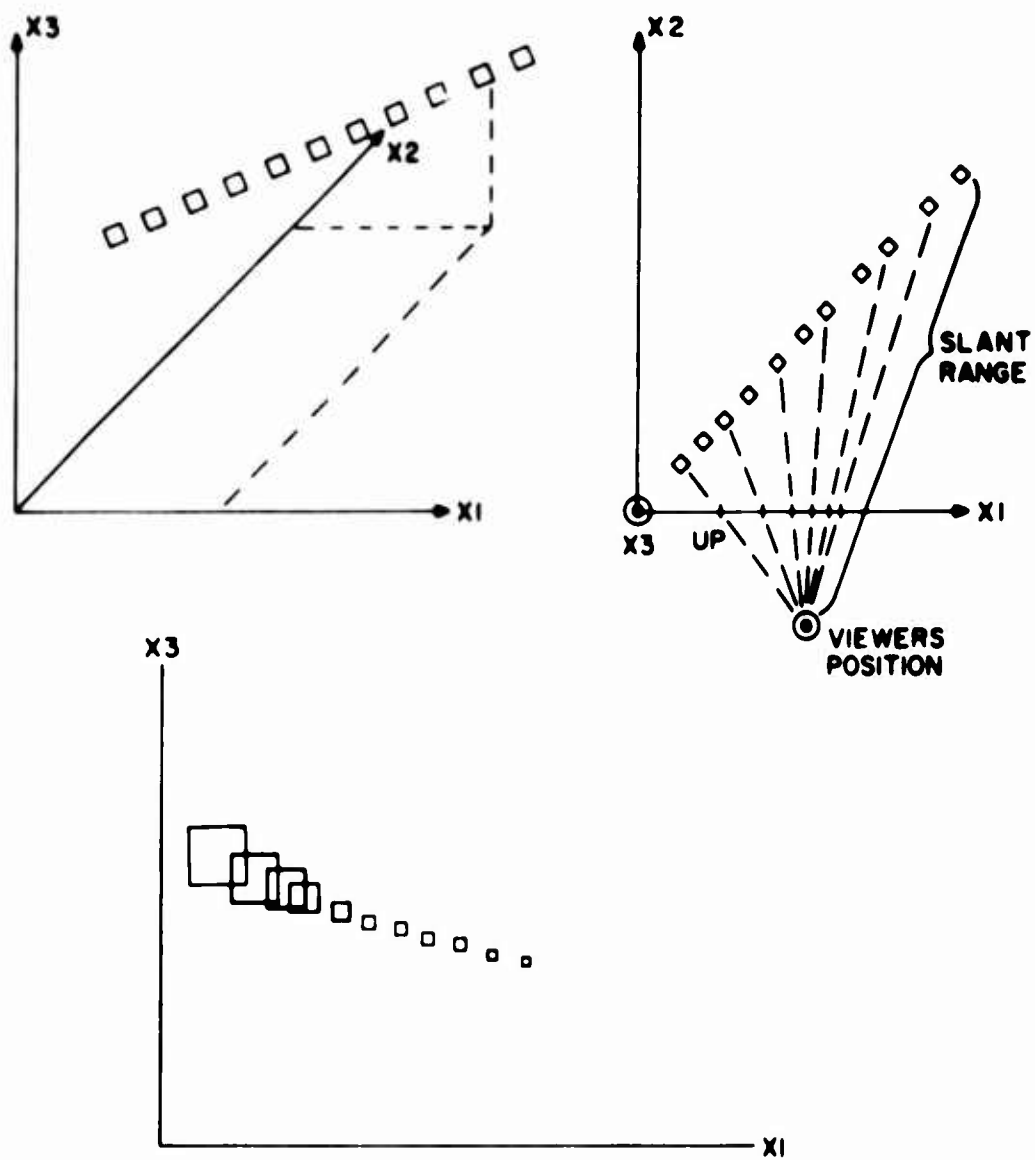


FIG. 7.19 PROJECTION PLOT



TA-658582-19

FIG. 7.20 PERSPECTIVE PLOT

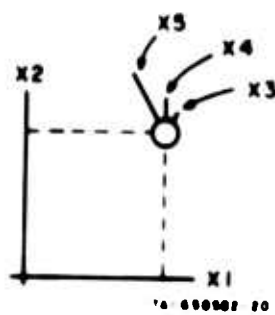


FIG 7.21 A METROGLYPH

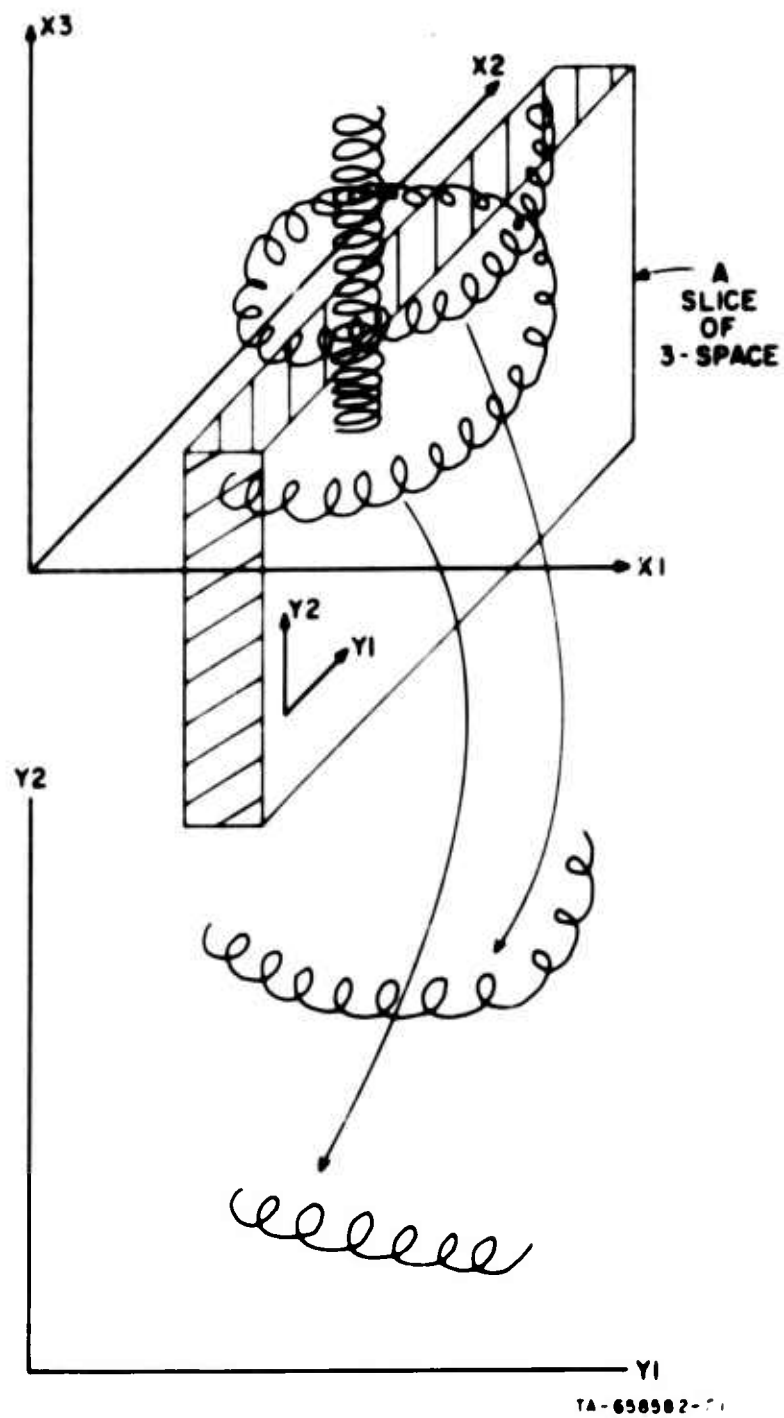


FIG. 7.22 A SLICE OF 3-SPACE

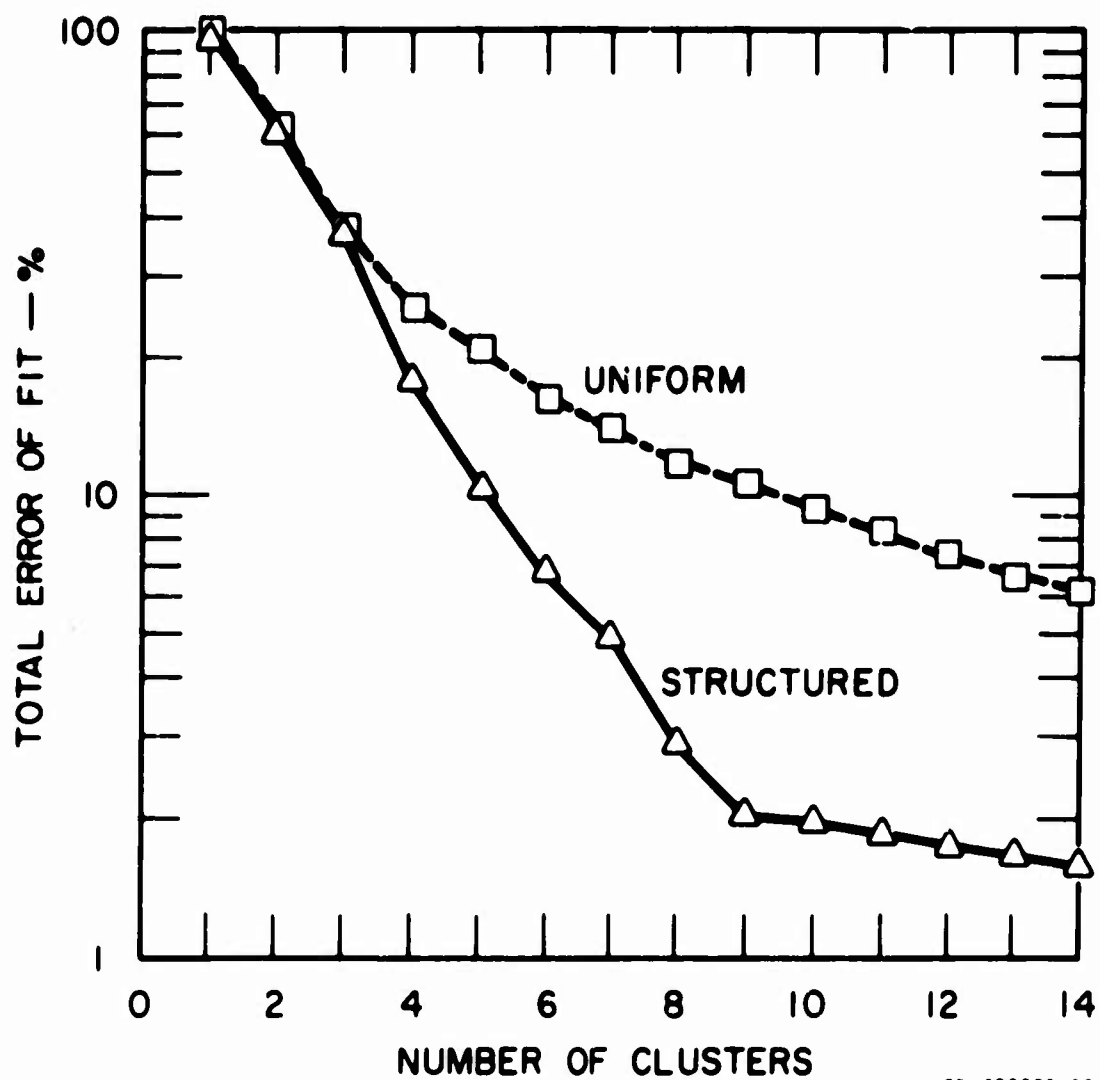
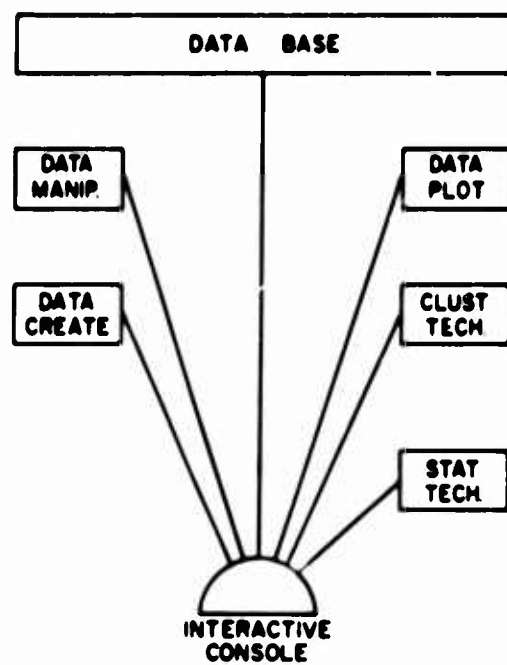


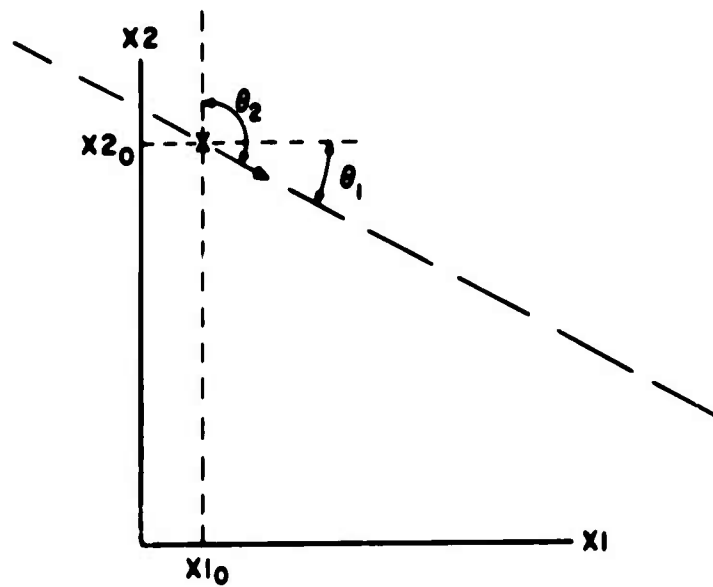
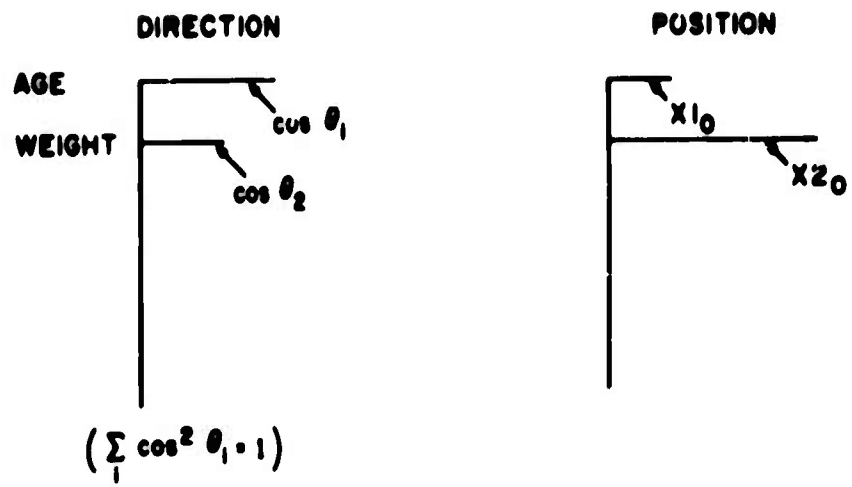
FIG. 7.23 CLUSTERING CHARACTERISTICS FOR UNIFORM AND STRUCTURED DATA



TA-000002-23

FIG. 7.24 INTERACTIVE DATA ANALYSIS SYSTEM





TA-650302-24

FIG. 7.25 SPECIFICATION OF A LINE IN 2 SPACE TO A COMPUTER

Rotating Principal Axes into Approximately Mutually  
Exclusive Categories  
Robert G. Ryder  
Child Research Branch  
National Institute of Mental Health

If one considers the history of personality research and thought over the past sixty years or so, there seems to be a general trend for terms that have originally designated categories to eventually become labels for dimensions. The category customarily becomes diminished in meaning to merely designate extreme scoring cases on the dimension. Hysterics become high scoring cases on a scale of hysteria, neurotics are those scoring high on neuroticism, extraverts score high in extraversion, and so on. At least where category or "type" means what Cattell calls a homostat, i.e., a collection of objects with similar attributes, this is likely to be the course of events for most typologies. Science tends to move toward greater precision. The idea of category is intrinsically binary, and hence usually involves throwing information away. Therefore, where situations permit, yes-no category concepts tend to drift toward continua, to permit more precise measurement potentialities.

Even the following progression is possible. An investigator factors a group of variables and obtains, say, seven dimensions. Each subject is then represented by a profile of seven factor scores. The investigator uses the factor scores as a basis for clustering individuals, and obtains 14 categories of subjects. In later work with this category system he becomes dissatisfied with the crudity of simply labeling a subject as in or out of a category, and so speaks, say, of the precise distance a subject is from the centroid of the category. But a subject has a distance from the centroid of each category so now each subject is represented by a profile of 14 distances. Thus science advances.

While it is convenient and useful to measure continua, there are often situational and conceptual restraints which lead to retaining binary distinctions. It is not customary for example, although perhaps it should be, to admit a student half-way to college, or to partly hire a person, or to assign a man to a position between two job categories, or to be semi-married, or to have a piece of furniture that is somewhere between a couch and a desk. In personality work, direct observation or conceptualization may conflict with the idea of a continuum. Many scores are possible on most measures of psychoticism; but a number of clinicians continue to maintain that being moderately psychotic is like being moderately pregnant.

In general, for personality work, the idea of a type or a syndrome is useful for those who must try to comprehend individual cases, or who want to put some flesh and blood reality into the psychometricians' abstractions. This way it is possible to imagine how various attributes may fit together as an organic unity, and to engage in some meaningful gestalt completion of a pattern that may be exhibited only in fragments. Since Mrs. Jones looks a bit like a classical hysteric in the way she presently acts, should we not keep an eye out for other parts of the pattern? Will she also exhibit conversion symptoms, la belle difference, or even fugue? A practical example of the use of the syndrome idea is to be found in Kennedy (1965). A decision is made as to whether a child fits one or another type of school phobia, and then the therapist makes explicit and direct use of his educated guesses as to various, so far unrevealed, aspects of the child's behavior.

There is no problem in using dimensions for measurement purposes while continuing to think in terms of categories, syndromes or types, as long as only one dimension is considered at a time. Types can simply refer to portions of the dimension's range. For example, a type might be derived by any means whatsoever, and individuals scaled on a dimension of distance from the centroid of the type, or perhaps according to the probability that they belong in the type, in which case the type label would refer to low distance scores or high probability values. There is, however, a practical problem where a number of different dimensions are used simultaneously. Since it is convenient to think of types as mutually exclusive, it is no longer possible simply to translate "type" into "extreme score." One deals instead with profiles of scores, and talks about a particular score profile as being such and such a type, with such and such a set of miscellaneous attributes. However, with even a moderate number of variables, the number of possible profiles becomes great, even if the number of popular profiles is not so great. Apart from the number of profiles, the situation is sloppy and inconvenient from the point of view of a human user. Why should a syndrome be defined in terms of three or five or seven variables if it can be defined in terms of one? The intent of the procedure to be described here is to reduce this sloppiness, to try and make the case with several variables similar to that with one variable. That is, in the ideal situation, no matter how many variables are used a category remains defined as an extreme score on one and only one dimension, even as categories remain mutually exclusive. To put it another way, the attempt is to juggle things in such a way that all, or as many as possible, of the score profiles are simple, single spike profiles.

The procedure is as follows. Take an  $N$  observations by  $T$  variables data matrix  $M$ , standardized for convenience in such a way that  $M'M$  is directly  $R$ , the  $T$  by  $T$  matrix of product moment correlations among variables. In general,

$$M = X \lambda' Y$$

where

$$X'X = I,$$

$$Y'Y = I$$

and  $\lambda$  is a diagonal matrix. Principal axes operations are used to obtain a matrix of loadings,  $Y \lambda'$ , and one of scores,  $X$ .

Customary procedure at this point would either be to leave the axes unrotated, or to rotate  $Y \lambda'$  in such a way as to yield simple structure among variables. The present suggestion is instead to rotate  $X$ , the matrix of factor scores, in such a way as to yield simple structure among subjects. Ideally, the resulting matrix, say  $XA$ , should be one where each subject has approximately zero scores on all dimensions but one, i.e., all profiles should be single spike in form.

In the examples to be presented, orthogonal rotation was employed, using normalized Guttman procedures and thus maximizing the likelihood that each observation will have as few as possible scores that are as extreme as possible.  $A$  is therefore a square orthonormal. Orthogonal rotation seems reasonable in view of the intent of this score or profile rotation (c.f. Ryder, 1964). That

is, there is no way each subject can have no more than one nonzero score without  $XA$  being orthogonal. Since

$$\begin{aligned}(XA)'(XA) &= A'X'XA \\ &= A'A \\ &= I\end{aligned}$$

this condition is fulfilled by orthogonal rotation.

It is possible for  $XA$  to be orthogonal while the columns of  $XA$  are not uncorrelated, as when scores for each dimension are either some positive value or zero. More categories are represented per dimension, however, if dimensions are bipolar, with positive score values, negative score values and zero score values, a situation that is guaranteed by the indicated standardization of  $M$ . If

$$\begin{aligned}\underline{1}' M &= \underline{0}' \\ \underline{1}' M Y \lambda^{-1} &= \underline{0}' \\ \underline{1}' X &= \underline{0}' \\ \underline{1}' XA &= \underline{0}'\end{aligned}$$

and

$$(XA)'(XA) = I$$

is directly the correlation matrix among profile rotated scores. If

$$\underline{1}' M \neq \underline{0}'$$

then possibly

$$\underline{1}' XA \neq \underline{0}'$$

but

$$(XA)'(XA) = I$$

still holds.

The rotated factor matrix  $F$  corresponding to the profile rotation  $XA$  is found by taking

$$\begin{aligned}F(XA)' &= M' \\ F &= M'XA \\ &= Y \lambda^2 X'XA \\ &= Y \lambda^2 A\end{aligned}$$

so that one can obtain  $F$  by either postmultiplying  $M'$  by  $XA$  or if  $A$  is known, postmultiplying  $Y \lambda^2$  by  $A$ . In either case the result is a matrix of factor loadings so rotated as to maximize the likelihood of simple factor score profiles.

As part of continuing research on the first years of marriage (Raush, Goodrich and Campbell, 1963; Goodrich and Ryder, 1966; Ryder and Flint, 1966; Ryder and Goodrich, 1966; Ryder, 1966), a great deal of information was gathered concerning a small group of suburban middle-class newlywed couples:  $N$  varied from 41 to 49 couples as a function of missing data. Since many more variables were measured than there were couples, an intensive effort was made to reduce the number of variables to a manageable size. What we did was to consider all the variables from a given technique or kind of technique, such as interview codes, questionnaires, or objective testing, and cluster them on a more or less ad hoc basis. The resulting clusters were then factored. The factors from these several techniques were then jointly factored in a final synthesis factoring which was based on only 15 variables (the 15 being previously extracted factors). This iterative factoring procedure also was intended to reduce the

likelihood of method factors, which customarily appear when data from several sources are jointly factored (Cartwright, Kirtner, and Fiske, 1963; Cartwright, and Roth, 1957; Forsythe and Fairweather, 1961; Gibson, Snyder, and Ray, 1955; Nichols and Beck, 1960).

Our factor data, and our factor score data, thus include three principal components based on an objective test called the Color Matching Test (CMT) (Goodrich and Boomer, 1963; Ryder and Goodrich, 1966; Ryder, 1966), four based on a content analysis of interview material, four based on ratings of interviews, and four based on the synthesis analysis. There were also four factors based on questionnaire material; but complications in their extraction and composition make it convenient to bypass them in the present discussion.

Factor scores for these various analyses were computed with and without profile rotation to get at least a rough idea of whether profile rotation increases the number of single spike profiles. In order to talk about spikes, it was necessary to define some convention as to what was an extreme score. It was decided to use that cutting point which would hypothetically permit perfect differentiation between two sets of scores. Procedure was as follows:

- 1) Two sets of factor scores, from the same principal components analysis, were tabulated in a frequency distribution of absolute scores, pooling between analyses and among axes.
- 2) The cutting point for "extreme" was located so as to leave (as closely as possible)  $N$  extreme scores overall, where  $N$  was as usual the number of observations.

If all profiles were either single spike or no spike, there could then be perfect, i.e., all single spike, representation of the scores using profile rotation, and no spiked profiles for the other set of scores (or vice versa).

Consider first the analysis of content codes from our interviews with the newlywed couples. Ten to 12 hours of interviewing per couple were subjected to a detailed content analysis, the data from which were summarized in 37 clusters of codings. Four principal components were computed both for the unrotated axes and for profile rotated data. The frequency distributions of numbers of spikes are given in Table 1 for unrotated and for profile rotated data.

It can be seen that there was a modest trend toward more single spike profiles with profile rotation. For the content analysis data alone, a check was made on the effects of conventional varimax rotation of factor loadings. Results for the corresponding factor scores, using the same cutting point, are given in Table 2.

Note that the number of single spike profiles is the same for unrotated data and data rotated in the usual manner.

Tables 3 and 4 compare unrotated and profile rotated scores for interview ratings and for the CMT, respectively. Notice that for CMT data the trend toward more single spike profiles with profile rotation is reduced to the vanishing point, and that the trend vanishes entirely for interview rating data.

The one remaining principal components analysis combines data from these other several sources, plus questionnaires, i.e., it is the synthesis analysis. Results for this are given in Table 5, and are a shade more encouraging.

On the assumption that there might be an interest in the qualitative changes that might derive from profile rotation, factor loadings for the synthesis analysis are given in Table 6. These are expressed not in terms of the first order factors that went into the synthesis analysis, but in terms of some of the variables on which those first order analyses were based. Variables included in Table 6 are those which loaded  $\geq .50$  on at least one factor from each rotation. The most striking difference between the two sets of factor loadings seems to be that profile rotation tends to bring the evaluative variables together in the same factor more than is the case for the unrotated factors.

The upshot of these various analyses is fairly disappointing. There is a slight tendency for profile rotation to increase the number of single spike profiles, at least for this sample; but so slight as to make it doubtful that a reasonable increment in single spike profiles is a dependable consequence of profile rotation. The trends have seemed so slight as to make it absurd to try and dignify them with inferential statistics. On the other hand these slight trends, combined with anomalies of these analyses (too small a sample and too many variables), are enough to keep alive the anticipation that with a larger sample and a cleaner set of variables the trends would prove to be nonchance and of a magnitude to make profile rotation worthwhile. It should be noted in passing that the total frequencies of spikes are determined by the procedure for setting cutting points for extreme scores. Juggling the cutting points around to ad hoc optima could lead to a much greater number of single spike profiles, and to a greater (or lesser) advantage for profile rotation. At any rate, data is now being collected on a far larger sample, and there should be more definitive information in due course.

## References

- Cartwright, D. S., Kirtner, W. L., & Fiske, D. W. Method factors in changes associated with psychotherapy. J. abnorm. soc. Psychol., 1962, 66, 164-175.
- Cartwright, D. S. & Roth, I. Success and satisfaction in psychotherapy. J. clin. Psychol., 1957, 13, 20-26.
- Forsythe, R. F. & Fairweather, G. W. Psychotherapeutic and other hospital treatment criteria: The dilemma. J. abnorm. soc. Psychol., 1961, 62, 598-604.
- Gibson, R. L., Snyder, W. U., & Ray, W. S. A factor analysis of measures of change following client-centered therapy. J. couns. Psychol., 1955, 2, 83-90.
- Goodrich, D. W. & Boomer, S. Experimental assessment of marital modes of conflict resolution. Family Process, 1963, 2, 15-24.
- Goodrich, D. W. & Ryder, R. G. Patterns of newlywed marriage. Paper given at the American Psychiatric Association meetings in May, 1966.
- Nichols, R. C. & Beck, K. W. Factors in psychotherapy change. J. consult. Psychol., 1960, 24, 388-399.
- Raush, H. L., Goodrich, D. W., & Campbell, J. D. Adaptation to the first years of marriage. Psychiatry, 1963, 26, 368-380.
- Ryder, R. G. & Flint, A. A. Vicissitudes of marital disputes: the Object Matching Test. Paper given at the American Orthopsychiatric Association meetings in April, 1966.
- Ryder, R. G. & Goodrich, D. W. Married couples responses to disagreement. Family Process, 1966, 5, 30-42.
- Ryder, R. G. Profile factor analysis and variable factor analysis. Psychol. Reports, 1964, 15, 119-127.
- Ryder, R. G. Two replications of Color Matching factors. Family Process, 1966, 5, 43-48.



Table 1

Unrotated and Profile Rotated Factor Scores  
Based on Interview Content Analysis

<u>Spikes per Profile</u>	<u>Unrotated</u>	<u>f</u>	<u>Quartimax Profile Rotated</u>
0)	29		24
1)	18		23
2)	2		2
3)	0		0
4)	0		0

Note: N = 49, four axes extracted from 37 variables.

Table 2

Scores corresponding to Varimax Rotation of Factor Loadings  
Based on Interview Content Analysis

<u>Spikes per Profile</u>	<u>f</u>
0)	26
1)	18
2)	4
3)	0
4)	0

Note: N = 48, four axes extracted from 37 variables.

Table 3

Unrotated and Profile Rotated Factor Scores  
Based on 21 Rated Interview Variables

<u>Spikes per Profile</u>	<u>Unrotated</u>	<u>f</u>	<u>Quartimax Profile Rotated</u>
0)	28		31
1)	17		15
2)	3		3
3)	1		0
4)	0		0

Note: N = 49, four axes extracted.



Table 4

Unrotated and Profile Rotated Factor Scores  
Based on 17 CMT Variables

<u>Spikes per Profile</u>	<u>Unrotated</u>	<u>f</u>	<u>Quartimax Profile Rotated</u>
0)	30		29
1)	16		18
2)	2		1
3)	0		0

Note: N = 48, three axes extracted.

Table 5

Unrotated and Profile Rotated Factor Scores  
Based on Synthesis Analysis

<u>Spikes per Profile</u>	<u>Unrotated</u>	<u>Quartimax Profile Rotated</u>
0)	31	27
1)	13	19
2)	4	1
3)	0	1
4)	0	0

Note: N = 48, four axes extracted from 15 variables  
that were themselves factors.

Table 6

Factors from the Synthesis Analysis <sup>a</sup>

Variable	Unrotated				Quartimax Profile Rotated			
	1	2	3	4	1	2	3	4
<u>Interview Code</u>								
Positive evaluation of housekeeping	-63	-12	-02	-12	13	-25	-54	23
Intersouse problems re friends	34	12	-19	54	27	-13	54	-28
W negative recall re home life	70	-05	-09	10	-27	14	55	-34
H high occupational ambition	-01	-03	60	-10	08	52	-28	-11
Difficulties with the W family	50	-24	10	30	-05		30	-54
<u>Interview Rating</u>								
H empathy	-24	-56	34	-49	-40	15	-72	-14
Overall evaluation of the marriage	-30	-56	19	-06	-08	-09	-58	-30
Couple (versus individual)								
Identification	-59	-19	28	12	35	-06	-59	-03
<u>GCT Score</u>								
H laughter	51	-36	03	13	-25	08	23	-54
<u>Questionnaire Variable</u>								
W involvement in job activities	-13	02	-50	26	14	-53	19	05
H report of contacts with H family	-29	55	07	21	51	07	09	40
H unhappiness or doubts re marriage	52	21	-31	-03	-24	-01	59	03
W unhappiness or doubts re marriage	66	17	-15	-13	-36	19	58	-04
W yields to H in disagreements	-44	-03	54	-03	30	30	-56	05

<sup>a</sup> Decimals omitted<sup>b</sup> Variables include only those loading  $\geq .50$ , on at least one factor of each rotation.

CLUSTER ANALYSIS AND THE SEARCH FOR STRUCTURE UNDERLYING  
INDIVIDUAL DIFFERENCES IN PSYCHOLOGICAL PHENOMENA \*

Ledyard R Tucker  
University of Illinois

Research on techniques for investigation of individual differences in psychological phenomena is related in several ways to the subject of this conference: cluster analysis of multivariate data. A first and major relation is to one of the important possible motivations for cluster analysis. This relation involves the general formulation of the research project at the University of Illinois on techniques for investigation of individual differences in psychological phenomena. Further relations involve some common technical problems and solutions.

Consideration of differences between individuals in psychological phenomena has had a long history dating back, undoubtedly, to the first thoughts of man in description of the behavior of other men. These considerations have continued and have entered the science of psychology at various points such as in the studies of the "personal equations" initiated by the astronomer Bessel, in the proposal of body types by Kretschmer, in the development of differential psychology as furthered by Galton. A variety of techniques have been developed for the study of the structure of individual differences in measurable attributes of individuals. Refinements and extensions of these techniques as well as the development of newer techniques are in progress for improving these studies. Even so, much of this work does not bear directly on some central problems in psychology and in relations between variables for single individuals. A first approximation in the description of this latter area is to describe it as a combination of traditional experimental psychology and differential psychology.

Cronbach, in his presidential address to the American Psychological Association in 1957, gave an excellent historical review and discussion of the contrast between experimental psychology and what he called correlational psychology which we may identify as differential psychology. He refers to the two disciplines of psychology as "two historic streams of method, thought, and affiliation which run through the last century of our science". He further noted that there has been recognition of the distinctions between these streams and that statements of hopes to bring them together have been made since the time of Wundt. For example, Cronbach stated that "Dashiell optimistically forecast a confluence of these streams, but that confluence is still in the making" and "Hull sought general laws just as did Wundt, but he added that organismic factors can and must be accounted for. He proposed to do this by changing the constants of his equation with each individual. This is a bold plan, but one which has not yet been implemented in even a limited way." A further comment by Cronbach which is quite relevant to this report is, "Tucker, though, has at least drawn blueprints of a method for deriving Hull's own individual parameters by factor analysis." I wish to add, hurriedly, that Cronbach is only partly correct in this reference to my work, which is not based on the Hullian learning curves; but, this work is concerned with the development of individual parameters as indicated by Cronbach. Further, this work includes the study

---

\* This paper is based on research jointly supported by the University of Illinois and the Office of Naval Research under contracts Nonr 134(39) and NC0014-66-C0010A03 .

of possible structural relations among parameters for individuals in a population. The confluence of experimental psychology and differential psychology is realized by an expansion of the concepts of psychological laws to involve individual parameters coupled with an extension of differential psychology to description of individual differences in these parameters, and thus in the psychological laws.

Study of parameters of a functional relation of a dependent variable to an independent variable, of which the study of learning functions is an example, is but one phase in the search for structure of individual differences underlying psychological phenomena. Work has progressed on developments in other areas of this general area of problems. Procedures have been investigated for description of individual differences in psychological scaling, both unidimensional scaling involving judgments in relation to named attributes such as preference or value, and multidimensional scaling. Closely related procedures have been investigated for the study of individual differences in judgments of similarity between pairs of stimuli, such as adjectives used to describe personality attributes. These techniques involve multivariate procedures closely related to factor analysis. A further development is the extension of factor analysis to consideration of data characterized by three modes of classification such as by individuals, traits measured, and occasion of measurement. Another example of three mode data would be the extent of reaction of individuals on several variables of reaction in several stimulus situations. The relation of pattern of reaction over variables for different stimulus conditions may be dependent on the individual who may be described by a group of parameters. Study of these individual parameters is involved in the search for structure of individual differences in psychological phenomena.

Some of the issues involved in the study of individual characteristics in psychological phenomena may be clarified by the following four attributes for description of scientific endeavors in psychology.

- A. Behaviors studied: single or multiple.
- B. Measures obtained: single or multiple.
- C. Values of the measures for individuals or within-individuals relations between measures for several variables.
- D. Individual differences: none, structured, chaotic.

The behaviors studied cover a wide range of activities of subjects, both in natural observational situations and experimental situations, such as conversation with other individuals, response to particular stimuli, performance on a given task, etc. The measures also cover a wide range of possibilities so that one or more measures may be obtained from any one behavior. For example, responses of subjects in a word association experiment may be measured by latency of response, galvanic skin reaction, and rareness of response word. Any one of these measures or several of them may be recorded for each response of a subject.

Attribute C is related to the common contrast made between S-R and R-R studies, but involves a basically distinct contrast. Many studies involve observation of the value of a single response measure from each of a number of behaviors for each subject and then study the relations between these measures over a group of subjects; these studies are classifiable as R-R studies and are examples of studies of the values of measures for individuals. It is

possible to obtain several response measures for each behavior of a subject in some category of behavior, such as a word association experiment, and to study the relation between the response measures over a number of behaviors for the same subject. These studies are classifiable as R-R studies but are studies of the within-individual relations between measures. Many studies classifiable as S-R studies involve within-individual relations between variables: a stimulus variable and a response variable. An extension of this class of studies may be denoted as  $S-(R_1, R_2)$  for which two measures of response are obtained for each value of a stimulus situation and the relations studied of these response variables to the stimulus variable and to each other.

Attribute D concerns the focus of experiments and assumptions made concerning differences between individuals in the phenomena being studied. Differential psychological studies emphasize the individual differences and tend to assume a structure in these differences. Many studies in experimental psychology minimize the differences between individuals, assuming either that there are no such differences or that the differences are chaotic and represent chance discrepancies from general laws of relations.

Using these attributes, many studies in differential psychology could be described as:

- A. multiple behaviors studied,
- B. single measure obtained for each behavior,
- C. value of each measure,
- D. structure of individual differences in these observations.

In contrast, many studies in experimental psychology could be described as:

- A. single behavior studied,
- B. multiple measures obtained,
- C. within-individual relations,
- D. assumption of no or chaotic individual differences.

A comparison of the search for individual differences in psychological phenomena with these two contrasting profiles is profitable. This project emphasizes:

- A. either single or multiple behaviors,
- B. multiple measures obtained for each behavior,
- C. within-individual relations,
- D. structure of individual differences in these relations.

This profile of attribute values has some similarity to each of the preceding profiles, but it is not a compromise between them. It goes beyond either of these types of studies and encompasses a number of very interesting and important problems. The motivation for this search for individual differences in psychological phenomena is not just to merge the two disciplines but is to solve problems not encompassed by either discipline.

A most interesting possibility in the structure of individual differences of within-individual relations between variables is that there exist clusters of individuals such that the within-individual relations are the same for all individuals in each such cluster and differ from cluster to cluster. If such clustering of individuals is the case, even within a reasonable approximation

to the actual structure of individual differences, the study of within-individual relations and the application of knowledge gained can be increased considerably in precision. Theories of learning, for example, could be constructed such that special cases would be applicable for each cluster of individuals. These special case learning theories would fit the learning behavior of individuals better than a learning theory that ignored individual differences. If there are clusters of individuals in the relations between abnormal psychological behavior and treatment, then the description of the effects of treatments could be increased in precision. Further, being able to place any mental patient within a cluster of individuals would aid materially with selection of treatments to lead to desirable behavior changes. It might be the case that seemingly conflicting theories of personality refer to different clusters of individuals in the dynamics of personality behavior and are special cases of a more general, but flexible, theory of personality which takes on different forms for the several clusters of individuals.

Before discussing cases of individual differences in within-individual relations some consideration will be given to work on personalizing regression estimation of criteria variables from selected predictor variables. Ghiselli (1956, 1960a) reported on work on the prediction of predictability in which he developed tests to predict the absolute values of errors of estimate in the regression of a criterion variable on a predictor variable. In terms of the errors of estimate, he could place individuals in two categories: one with low absolute errors of estimate and the other with high errors of estimate. By constructing a new measure using item analysis procedures he was able to approximate the placement of individuals in these classes. This constitutes a simple case of categorizing individuals as to the relation between a criterion variable and a predictor variable.

In a second activity, Ghiselli (1960b) worked on the differentiation between tests as to the accuracy with which they predict a criterion for a given individual. In this case two predictor variables were considered separately and errors of estimation were obtained for each predictor in a regression with the criterion variable. The absolute values of these errors of prediction were used and categories were established according to which error of prediction was larger in absolute value. Again, a new measure was constructed by item analysis procedures to approximate the placement of individuals in the categories. This is a most interesting possibility for the categorization of individuals according to the relations between variables. A point to note is that the individuals do not form homogeneous groups as to either the predictor variables or the criterion variables. The categories are related to the relations between the criterion variable and the predictor variables.

The work by Ghiselli is related to the study by Frederiksen and Melville (1954) on differential predictability of test scores and to Saunders' work (1955, 1956) on moderator variables. More recently, Cleary (1966) has proposed a technique for investigation of the possibility of developing systems of individualized regression weights in estimation of a battery of criteria from a battery of predictor variables. Given scores  $x_{pj}$  of persons  $p = 1, 2, 3, \dots, P$  on predictor variables  $j = 1, 2, 3, \dots, J$  and  $y_{pk}$  of the persons on criterion variables  $k = 1, 2, 3, \dots, K$  the personalized linear regression equation can be written as

$$1) \quad y_{pk} = \sum_j w_{pjk} x_{pj} + e_{pk}$$

where  $w_{pjk}$  are the personalized regression weights and  $e_{pk}$  are the errors of estimate. The personalized regression weights are defined by

$$2) \quad w_{pjk} = \sum_m b_{pm} a_{mjk}$$

for dimensions  $m = 1, 2, 3, \dots, M$  of a regression weight space and where  $b_{pm}$  are coefficients for the persons and  $a_{mjk}$  are coefficients for combinations of predictors  $j$  and criteria  $k$ . This system degenerates to the usual regression system when there is one dimension and all  $b_{pm}$  are unity.

Otherwise, this system provides for individual differences in the regression weights within the limits of the number of dimensions utilized. To obtain non-trivial solutions the number of dimensions must be fewer than the number of criterion variables. Note that determination of the coefficients in this system depends only on knowledge of the predictor and criterion variable scores. In an experimental application of this system to a case involving five criteria, two batteries of five predictors, and two samples of individuals Cleary found that the use of two dimensions in the regression weight space for each battery of predictors markedly reduced the sum of squared errors of estimate and that the  $a_{mjk}$  coefficients were very stable when determined

separately from the two samples. The person coefficients  $b_{pm}$  had one stable dimension when determined for each sample separately from the two batteries of predictors. While the person coefficients  $b_{pm}$  are determined in this model from knowledge of both the predictor and criterion scores, which makes use of the model questionable in applied situations, approximations to these coefficients may be obtained from measures developed by test construction and item analysis techniques. A very interesting possibility is that the individuals might be distributed in a number of clusters according to the values of their coefficients  $b_{pm}$ . If this were the case, categories might be established such that different regression systems were appropriate for the different categories. Such categorization would be extremely important in that it would indicate the existence of sub-populations of people for which different laws of relations existed between measures of behavior. Knowledge of these differences in laws of relation would add materially to our knowledge of psychological phenomena.

Extensive and critical studies should be conducted as to the possibility of the clustering of individuals as to within-individual relations between variables. For an example of such studies consider the area of color perception. Illustrative data for such a study is given in Table 1. These data are fictitious, being constructed to present a simpler version of results obtained by Helm and Tucker (1962); these data, however, represent fairly faithfully some of the major aspects of the results obtained by Helm and Tucker from real data. In the real data, Helm obtained measures of judged interpoint distances between stimulus objects for each pair of such objects using the method of triad-ratio judgments. These measures were obtained separately for each subject. The stimulus objects used by Helm were ten hexagonal tiles, 2 inches across, each painted with a different color, such that the ten colors were of the same lightness and saturation and formed an equally spaced circle of hues. The data in Table 1 are interpoint distances for pairs of eight color stimuli and six individual subjects. A major attribute of the fictitious data in Table 1 is that it is constructed so that the model for studying individual differences in



multidimensional scaling (Tucker and Messick, 1963) fits perfectly. One of the technical problems to be discussed subsequently in this paper concerns the fitting of the model to real data involving discrepancies of the model from the observed data. Individuals 1, 2, and 3 have normal color vision while individuals 4, 5, and 6 have progressively weaker color vision. This distribution of subjects as to color vision has relatively too few individuals with normal color vision but seems to represent fairly well the progression of color weak subjects as appearing in the Helm and Tucker results. There is an unresolved question as to the distribution of relative extents of weakness in color vision in the population.

Let us compare the color judgment data of the Helm and Tucker study with the four attributes of studies discussed earlier. A series of behaviors exists for each subject: the judgments of relative differences of pairs of stimuli in triads of stimuli. Two measures are obtained for each behavior: the judged ratios of relative differences between stimuli in the two less different pairs of stimuli to the relative differences between stimuli in the most different pair in a triad of stimuli. These data have been analysed for each subject to relative interpoint distances between stimuli in each pair of stimuli from the set of stimuli used in the study. These relative interpoint distances could be analysed for each subject to uncover a multidimensional scaling of the perceptual space for that subject, a step that was performed by Helm. These multidimensional scalings constitute within-individual relations among the measures. Analysis of the matrix of relative interpoint distances, such as in Table 1, by the Tucker and Messick model for individual differences in multidimensional scaling is an investigation of the structure of individual differences in these within-individual relations. Thus, this study illustrates the profile of attribute values for the search for individual differences in psychological phenomena.

Analysis of the individual interpoint distances for the structure of the individual differences takes these interpoint distances as input data and forms a matrix, which is designated  $X$  and is illustrated in Table 1, with a row for each pair of stimuli and a column for each individual. This analysis proceeds to determine what is called here the characteristic components of the matrix  $X$  by a technique based on the theorem by Carl Eckart and Gale Young (1936) on the approximation of a matrix by another of lower rank. This technique is closely related to the method of principal components proposed by Harold Hotelling (1933). For the sake of clarity, discussion of several technical points is being postponed to following presentation of the analysis technique as applied to the data in Table 1. Steps in the analysis are outlined below.

A. Compute the matrix product  $X'X$  which contains the sums of squares of entries in the columns of  $X$  as diagonal entries and sums of products of pairs of entries in each pair of columns as off-diagonal entries.

B. Determine a scaling constant  $k^2$  by dividing the number of individuals by the sum of the diagonal entries in  $X'X$  (this sum equals the sum of squares of the entries in  $X$ ).

C. Multiply  $X'X$  by the scaling constant  $k^2$ .

D. Determine the characteristic roots and unit length vectors of  $k^2X'X$ . Let the roots be designated by  $\gamma_m^2$  and be arranged in descending order, and let the corresponding vectors be designated by  $V_m$ . The roots for the illustrative example are listed at the left of Table 2.

E. Select the  $r$  largest roots (a point to be discussed) and form the matrix  $Z$  containing as row vectors  $\gamma_m V_m$  for the selected roots. The matrix  $Z$ , in transposed form, for the example is given in the middle of Table 2. The



entries in this matrix are the scores of the individuals on the selected characteristic components.

F. Determine the matrix  $B$  of loadings of stimulus pairs on characteristic components by

$$3) \quad B = XZ'(ZZ')^{-1}.$$

Each column vector,  $B_m$  may be determined by

$$4) \quad B_m = XZ'_m Y_m^{-2} = X V_m Y_m^{-1}$$

where  $Z_m$  is the  $m$ 'th row vector of  $Z$  and  $V_m$  is the  $m$ 'th characteristic vector written as a column vector. The matrix  $B$  for the example is at the right of Table 2.

G. Construct an  $r$  dimensional space corresponding to the matrix  $Z$  in which each individual,  $i$ , is represented by a point having coordinates  $z_{mi}$  on the  $m$  orthogonal, coordinate axes. For the example, this space is two dimensional and is presented in Figure 1 with a solid point for each of the individuals. The configuration of points in this space represents the structure of the individual characteristics underlying the psychological phenomenon of color vision as this phenomenon is reflected in the judgments made for the selected group of stimuli.

H. Inspect the configuration of points in the space constructed in step G for interesting characteristics such as clusters of individuals. In the example, the three normal subjects are colinear from the origin and can be thought of as constituting a cluster. The three individuals having weakness in color perception do not form a cluster but have points located at varying distances from the cluster of points for individuals having normal color vision. These distances correspond to the extent of deficiency in color vision of these individuals. Results of this inspection may be described, in part, by selection of points in the space that may be considered as conceptual, or idealized individuals which represent interesting locations in the configuration of points for the actual individuals. In the example, two idealized individuals were selected: one to represent the individual having normal color vision and the other beyond the most severely color-weak observed individual so as to, possibly, represent an individual who is totally color blind. These are idealized individuals A and B and are indicated in Figure 1 by open circles.

I. Construct a matrix  $Z^*$  of scores of idealized individuals on characteristic components containing the coordinates of the points for the selected idealized individuals. This matrix has a column for each idealized individual and a row for each characteristic component. For the example, the matrix  $Z^*$ , in transpose form, is given at the left of Table 3.

J. Determine the matrix  $X^*$  of interpoint distances between pairs of stimuli for the idealized individuals by

$$5) \quad X^* = BZ^*.$$

This matrix for the example is given on the right of Table 3.

K. Using the interpoint distances in each column of  $X^*$ , separately by column, perform a multidimensional scaling to obtain the perceptual space for each idealized individual. These two spaces resulting from the multidimensional scaling for the two idealized individuals in the example are given in Figure 2. The space for idealized individual A which represents normal

color vision is two dimensional and has a circular configuration of points for the selected color stimuli. This is the expected result. The perceptual space for idealized individual B is also two dimensional, but has the stimuli in a semi-circle such that would be produced by folding the circle for normal color vision on an axis from Y to B. This was an unexpected result found by Helm in his multidimensional scaling of individual interpoint distances and which appeared in the analysis by Helm and Tucker. The expectation was that the loss of color vision would result in a one dimensional perceptual space. This appears not to be the case. These results raise several interesting conjectures which could be investigated experimentally as to the perception of color by color-blind individuals. However, discussion of these conjectures here would take us too far afield from the major theme of this paper.

The format of analysis outlined in the preceding paragraphs is of quite wide applicability for a variety of types of data. One important requirement is that the data for each subject be measures of a single dependent variable for various values of independent variables. In the preceding example, the independent variable was the set of colored stimulus pairs formed by the Cartesian product of the set of colored stimuli with itself, excluding identical pairs. The dependent variable was the judged interpoint distances. In the generalized format, the observations of the dependent variable for each individual would be recorded in a column of the matrix  $X$ . Each row of  $X$  would be for some particular values of the independent variables. Steps A through J would be conducted as described while step K would be altered to a form appropriate to study of the relation of the dependent variable to the independent variables. For another example consider a study of the learning of some task. The independent variable would be the series of trials or learning periods. The dependent variable would be a measure of the performance of a subject on each trial. There would be a row of matrix  $X$  for each trial and a column for each individual with entries being the measures of performance. The matrix  $X^*$  obtained in step J would contain measures of performance for idealized individuals on the trials so that the series of entries in each column of  $X^*$  could be used to develop a learning curve for the corresponding idealized individual. Another example could be the study of preferences among pairs of stimuli for which the dependent variable was ratings of relative preference. Results could yield a preference scale for each idealized individual. Still another example could involve semantic differential ratings of concepts on bipolar adjective scales. Each row of matrix  $X$  would be for a pairing of a concept with a bipolar scale. Step K would involve the determination of the connotative semantic space for each idealized individual.

There are several technical matters involved in the analysis which warrant consideration in this report. The measures of the dependent variable should be such as to support a study of the relation of the dependent variable to the independent variables for each individual. Further, these measurements for the dependent variable should be interpretable as on either an interval scale or a ratio scale for each individual. In case the measures are on an interval scale with a meaningless origin, the origin for each individual should be set at the mean for the individual so that deviations from the mean for the individual are used in the analysis. This step converts the interval scale measurements to a type of ratio scale measurements of discrepancies from the individual mean. Necessity for a ratio class of measurement lies in the model underlying the analysis of steps A through J.

Characteristic component analysis as used here is related to factor analysis, especially to obverse factor analysis for which factors among people are

determined. There are, however, several important distinct features. Consider the statistical model for regular factor analysis as given in equations (6) and (6').

$$6) \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1r} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2r} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nr} \end{pmatrix} \begin{pmatrix} u_1 & 0 & 0 & \dots & 0 \\ 0 & u_2 & 0 & \dots & 0 \\ 0 & 0 & u_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_n \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_r \\ \hline u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}$$

$$6') \quad \underline{x} = (A \mid U) \begin{pmatrix} \underline{\xi} \\ \underline{u} \end{pmatrix}$$

For each individual sampled from some population there is a random variable vector  $\begin{pmatrix} \underline{\xi} \\ \underline{u} \end{pmatrix}$  of dimensionality  $(r + n)$  where there are  $r$  common factors and  $n$  unique factors. Entries in this vector are the factor scores for the individual. The factor matrix  $(A \mid U)$  is a transformation on the factor score vector to yield the random variable vector  $\underline{x}$  of observed scores on the  $n$  variables. In the population, the random variable vector  $\begin{pmatrix} \underline{\xi} \\ \underline{u} \end{pmatrix}$  of factor scores has a density function with mean vector  $\begin{pmatrix} \underline{\psi} \\ \underline{\omega} \end{pmatrix}$  and a covariance matrix  $\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$ . The density function for the random variable  $\underline{x}$  of observed scores has a mean vector of  $\underline{\mu}$  and covariance matrix  $\Sigma$ . The relation of the mean vector for observed scores to the mean vector for factor scores is given in equation (7)

$$7) \quad \underline{\mu} = (A \mid U) \begin{pmatrix} \underline{\psi} \\ \underline{\omega} \end{pmatrix},$$

and the relation of the covariance matrix of the observed scores to the covariance matrix of factor scores is given in equation (8)

$$8) \quad \Sigma = (A \mid U) \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A' \\ U' \end{pmatrix}.$$

From a random sample of individuals, the vector of sample means  $\bar{x}$  is an estimate of the vector  $\mu$ ; and the covariance matrix  $S$  is an estimate of  $\Sigma$ . Factor analytic methods produce estimates  $\hat{A}$  and  $\hat{U}$  of the population transformation parameters  $A$  and  $U$ . A second set of methods are used to obtain estimates of the factor scores; these methods include such procedures as regression estimates and Bartlett's (1936) procedure to minimize the uniquenesses. Guttman (1955) has pointed out the insolubility of the factor score problem.

Obverse factor analysis interchanges the role of the individual and the attribute measured in the model for factor analysis. Any attribute measured is considered as sampled from a population of attributes and is associated with random variable vectors having entries for the individuals. In this case the individuals are taken as fixed so that the factor scores of the individuals are parameters of the model and are estimated by the analysis. For this case, the loadings of the variables are inaccessible in the same sense as the factor scores were inaccessible.

Since the search for individual differences in psychological phenomena requires both the loadings of the attributes measured and the factor scores, neither the direct factor analysis model nor the obverse model is appropriate. A third model is employed.

The two major aspects of the factor analytic model for the present discussion are the assumption of unique factors and the assumption that the individuals or attributes measured are sample elements from a population. In contrast, the model for characteristic component analysis does not include unique factors and assumes that the individuals and attributes measured are the equivalent of fixed effects. The model for this analysis is given in equation (9)

$$9) \quad x_{ji} = \sum_{m=1}^r b_{jm} z_{mi} + \epsilon_{ji}$$

where  $\epsilon$  is a random variable with mean = 0 and standard deviation =  $\sigma_{ji}$ , and is independent for each  $ji$  combination. Exclusion of the unique factor aspect of the factor analytic model implies that the group of variables,  $j$ , cover the domain so that particular variables are not dependent on specific influences. Postulation of individuals as fixed effects is necessary to enable estimation of both the  $b_{jm}$ 's and the  $z_{mi}$ 's as parameters of the model. Such estimates are needed for the complete procedure involving the individual space, selection of idealized individuals, and estimation of observations for these idealized individuals. The procedure described provides a least squares fitting of the model to the data. Also, in case the  $\sigma_{ji}$  are constant for all  $ji$  combinations, the procedure is a maximum likelihood estimation as indicated by Young (1941). In case the  $\sigma_{ji} = \alpha_j \beta_i$  where the  $\alpha_j$  are known for the variables and  $\beta_i$  are known for the individuals, a slightly more complex procedure should be used. As noted by Anderson and Rubin (1956), it is necessary that the  $\alpha_j$  and  $\beta_i$  be known.

Determination of the number  $r$  of dimensions to be used in the analysis for any particular body of data is an unsolved problem quite analogous to the number of factors problem. Several properties of the series of roots  $\gamma_m^2$  may be noted in this context as yielding some guides to the selection of the number of dimensions to be used. First, all roots are non-negative. Second, the sum

of squares of the errors of approximation of the data from the model for any given number of dimensions chosen is given by the sum of the remaining roots. Third, the sum of squares of the approximations to the data is given by the sum of the roots for the dimensions used in the approximation. Thus, one possibility is to use as many dimensions as necessary to obtain some desired degree of goodness of fit of the model to the data. Cumulative sums of the roots will aid in determining the number of dimensions necessary. A second possibility is to inspect the series of roots for some break in the relation between root size and root number. For this criterion, one postulates that the individual space would be of some small dimensionality except for the discrepancies involved in making the observations. Then, there should be two laws of formation for the series of roots, one for the dimensions relevant to the individual space and a second for the discrepancies. If this be the case, there should be a break in the relation between root size and root number. Such changes in form of relation have been observed. Further, for some bodies of data, there appeared to be a linear relation between root size and root number for a large number of roots beyond a small number of initial roots which were larger than would be expected from this linear relation. At present, this inspection of the series of roots seems to be the best procedure available.

A single dependent variable has been involved in the discussion to this point; however, the extension of factor analysis to a model for three mode factor analysis permits the investigation of cases when measures are made on several dependent variables for each pattern of values of the independent variables. For an example of this class of data consider the complex tracking task investigated by Parker and Fleishman (1960). The subject was to control a dot on a cathode ray oscillograph using a control stick and rudder pedals as in a standard aircraft control system. Movement of the dot was introduced electronically and the subject's task was to keep the dot centered on the oscillograph as well as to avoid sideslip which was indicated by a separate meter. Measures were obtained of horizontal error, vertical error, sideslip error, and time-on-target for each of a number of stages of practice. The study by Parker and Fleishman involved 203 individuals, 10 stages of practice, and the four dependent variables listed above. These may be interpreted as the three modes for identification of the data as defined by Tucker (1964, 1966). The model for three mode factor analysis that would be appropriate in the present context would be an extension from equation (9) (present use of letters is not to be confused with previous use):

$$10) \quad x_{ijk} = \sum_{m=1}^M \sum_{p=1}^P \sum_{q=1}^Q a_{im} b_{jp} c_{kq} g_{mpq} + \epsilon_{ijk} .$$

The observed data are denoted by  $x_{ijk}$  which are entries in a three mode matrix  $X$  with rows for individuals, columns for stages of practice, and strata for dependent variable. Parameters of the model are contained in the two mode matrices  $A$ ,  $B$ , and  $C$  plus the three mode matrix  $G$ . The matrix  $A$  has rows for observed individuals and columns for idealized individuals; matrix  $B$  has rows for observed stages of practice and columns for idealized stages of practice; matrix  $C$  has rows for observed dependent variables and columns for idealized dependent variables. Matrix  $G$  is called the core matrix and has measures of the idealized dependent variables at the idealized stages of practice for the idealized individuals. Again,  $\epsilon$  is a random variable with mean = 0 and standard deviation =  $\sigma_{ijk}$ . Analysis methods are con-

siderable extensions of the type outlined earlier. The major aspect of this analysis especially relevant for this conference involves the study of the matrix  $A$  for clusters of individuals. If such clusters exist, then the relations of the dependent variables on the independent variables would be the same for individuals in each cluster and different for individuals from different clusters.

Major emphasis has been placed on the categorization of individuals as to the relations between variables. To attempt to establish categories according to the values of measures such that all individuals within a category can be considered as replicates and for which there would be the same expectation as to other measures seems to be an unrealistic and hopeless endeavor. If categories can be established as to the relations between measures, the individuals within a category could differ while the same dynamic laws applied. These categories would be especially interesting and relevant in scientific knowledge as well as in application to many problems.

#### References

- Anderson, T. W. and Rubin, H. Statistical inference in factor analysis. Proceedings of the third Berkeley symposium on mathematical statistics and probability, Volume V. Berkeley, California: University of California Press, 1956. Pages 111 - 150.
- Bartlett, M. S. The statistical conception of mental factors. Brit. J. Psychol., 1937, 28, 97-104.
- Cleary, T. Anne. An individual differences model for multiple regression. Psychometrika, 1966, 31, 215-224.
- Cronbach, Lee J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.
- Eckart, Carl and Young, Gale. The approximation of one matrix by another of lower rank. Psychometrika, 1936, 1, 211-218.
- Frederiksen, N. and Melville, S. D. Differential predictability in the use of test scores. Educ. psychol. Measmt., 1954, 14, 647-656.
- Ghiselli, E. E. Differentiation of individuals in terms of their predictability. J. appl. Psychol., 1956, 40, 374-377.
- Ghiselli, E. E. The prediction of predictability. Educ. psychol. Measmt., 1960a, 20, 3-8.
- Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. Educ. psychol. Measmt., 1960b, 20, 675-684.
- Guttman, L. The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. Brit. J. Stat. Psychol., 1955, 8, 65-81.

- Helm, C. E. and Tucker, L. R. Individual differences in the structure of color perception. Am. J. Psychol., 1962, 75, 437-444.
- Hotelling, H. Analysis of a complex of statistical variables into principal components. J. educ. Psychol., 1933, 24, 417-441, 498-520.
- Parker, James F., Jr. and Fleishman, Edwin A. Ability factors and component performance measures as predictors of complex tracking behavior. Psychol. Mono., 1960, 74, No. 16.
- Saunders, D. R. The moderator variable as a useful tool in prediction. Proceedings of the 1954 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1955, pp. 54-58.
- Saunders, D. R. Moderator variables in prediction. Educ. psychol. Measmt., 1956, 16, 209-222.
- Tucker, L. R. Determination of parameters of a functional relation by factor analysis. Psychometrika, 1958, 23, 19-23.
- Tucker, L. R. and Messick, S. An individual differences model for multidimensional scaling. Psychometrika, 1963, 28, 333-367.
- Tucker, L. R. The extension of factor analysis to three-dimensional matrices. In W. Frederiksen and H. Gulliksen (eds.), Contributions to mathematical psychology. New York: Holt, Rinehart and Winston, 1964.
- Tucker, L. R. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-311.
- Tucker, L. R. Learning theory and multivariate experiment: illustration by determination of generalized learning curves. In R. B. Cattell (ed.), Handbook of multivariate experimental psychology. Chicago: Rand McNally, in press.
- Young, G. Maximum likelihood estimation and factor analysis. Psychometrika, 1941, 6, 49-53.



Table 1  
Matrix X of Individual Interpoint Distances  
Fictitious Color Judgment Data\*

Stimulus Pairs	Individuals					
	1	2	3	4	5	6
a-b	8	7	8	8	10	8
a-c	14	13	16	16	18	16
a-d	18	17	20	18	19	11
a-e	20	18	22	18	16	4
a-f	18	17	20	18	19	11
a-g	14	13	16	16	18	16
a-h	8	7	8	8	10	8
b-c	8	7	8	8	10	8
b-d	14	13	16	13	11	3
b-e	18	17	20	18	19	11
b-f	20	18	22	21	23	17
b-g	18	17	20	20	24	20
b-h	14	13	16	16	18	16
c-d	8	7	8	8	10	8
c-e	14	13	16	16	18	16
c-f	18	17	20	20	24	20
c-g	20	18	22	22	26	22
c-h	18	17	20	20	24	20
d-e	8	7	8	8	10	8
d-f	14	13	16	16	18	16
d-g	18	17	20	20	24	20
d-h	20	18	22	21	23	17
e-f	8	7	8	8	10	8
e-g	14	13	16	16	18	16
e-h	18	17	20	18	19	11
f-g	8	7	8	8	10	8
f-h	14	13	16	13	11	3
g-h	8	7	8	8	10	8

Stimulus colors: a - Red                      e - Green  
                          b - Orange                      f - Blue-Green  
                          c - Yellow                      g - Blue  
                          d - Yellow-Green                      h - Purple

\* Designed in accord with results reported by Helm and Tucker (1962).



Table 2  
Characteristic Components Analysis of Matrix X

Characteristic Roots of $(k^2 X'X)^*$		Scores of Individuals on Characteristic Components			Loadings of Stimulus Pairs on Characteristic Components		
		Individual	I	II	Stimulus Pair	I	II
I	5.915	1	.97	-.09	a-b	8.4	5.6
II	.085	2	.88	-.08	a-c	15.5	10.3
III	0.000	3	1.07	-.10	a-d	17.4	-17.6
IV	0.000	4	1.02	-.02	a-e	16.7	-42.7
V	0.000	5	1.14	.07	a-f	17.4	-17.6
VI	0.000	6	.85	.24	a-g	15.5	10.3
					a-h	8.4	5.6
					b-c	8.4	5.6
					b-d	11.8	-30.2
					b-e	17.4	-17.6
					b-f	20.4	- 2.2
					b-g	20.2	13.5
					b-h	15.5	10.3
					c-d	8.4	5.6
					c-e	15.5	10.3
					c-f	20.2	13.5
					c-g	21.9	14.6
					c-h	20.2	13.5
					d-e	8.4	5.6
					d-f	15.5	10.3
					d-g	20.2	13.5
					d-h	20.4	- 2.2
					e-f	8.4	5.6
					e-g	15.5	10.3
					e-h	17.4	-17.6
					f-g	8.4	5.6
					f-h	11.8	-30.2
					g-h	8.4	5.6

\*  $k^2 = N / \text{Trace } (X'X)$

Table 3

## Interpoint Distances for Idealized Individuals

Scores of Idealized Individuals on Characteristic Components			Interpoint Distances for Idealized Individuals		
<u>Idealized Individual</u>	<u>I</u>	<u>II</u>	<u>Stimulus Pairs</u>	<u>A</u>	<u>B</u>
A	.97	-.09	a-b	8	8
B	.73	.28	a-c	14	14
			a-d	18	8
			a-e	20	0
			a-f	18	8
			a-g	14	14
			a-h	8	8
			b-c	8	8
			b-d	14	0
			b-e	18	8
			b-f	20	14
			b-g	18	18
			b-h	14	14
			c-d	8	8
			c-e	14	14
			c-f	18	18
			c-g	20	20
			c-h	18	18
			d-e	8	8
			d-f	14	14
			d-g	18	18
			d-h	20	14
			e-f	8	8
			e-g	14	14
			e-h	18	8
			f-g	8	8
			f-h	14	0
			g-h	8	8

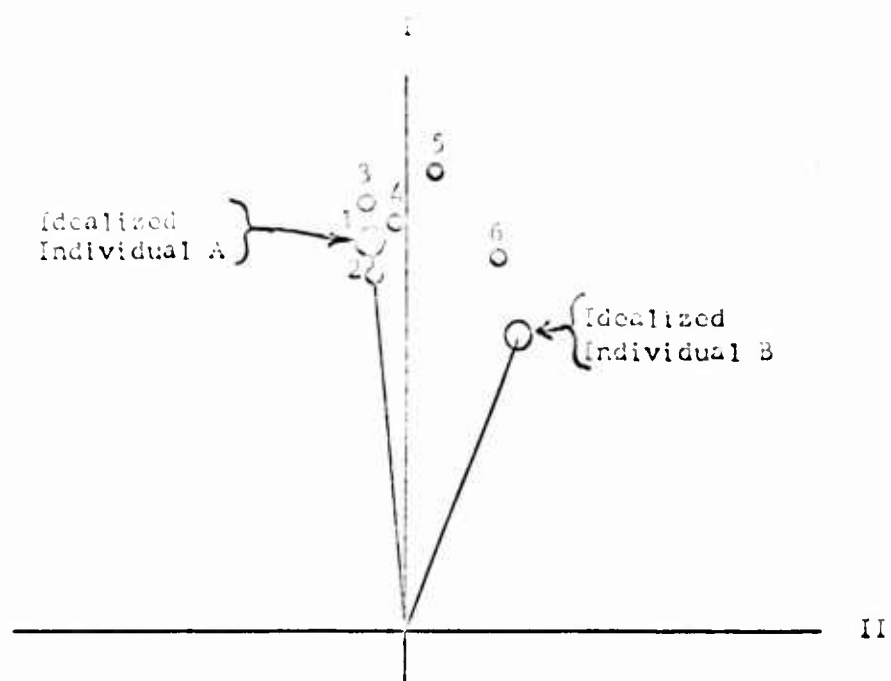


Figure 1  
Characteristic Component Space for Individuals

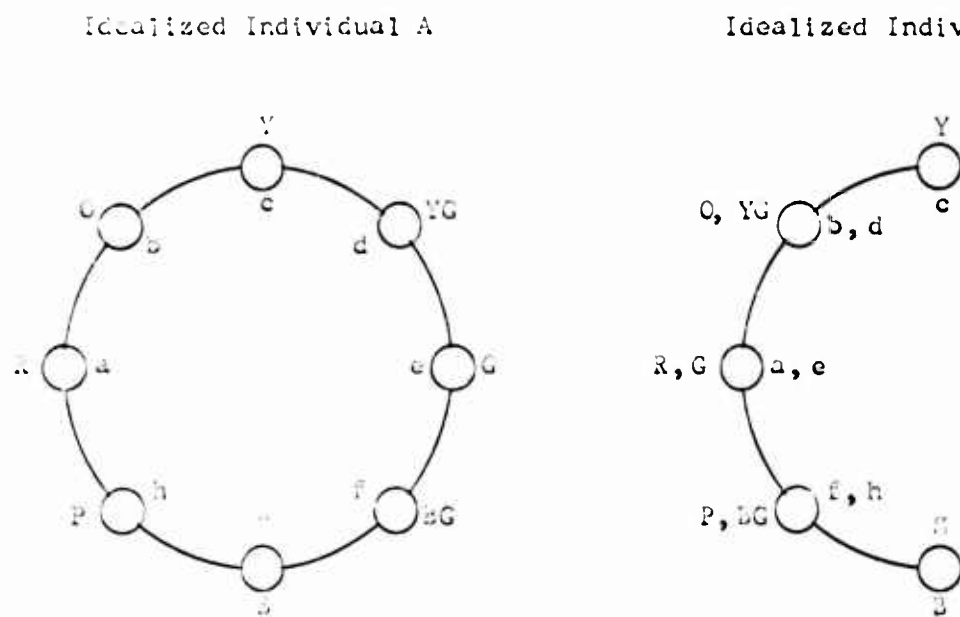


Figure 2  
Color Perceptual Spaces for Idealized Individuals

## THE MAXOF CLUSTERING MODEL<sup>1</sup>

Raymond E. Christal  
Air Force Personnel Research Laboratory

and

Joe H. Ward, Jr.  
Southwest Educational Development Laboratory

This paper describes a highly flexible technique for clustering people or things into categories. If a complete hierarchical structure is desired, such as the establishment of a biological taxonomy, then the MAXOF Clustering Model will yield an optimum solution in terms of a criterion established by the investigator. If the desire is to cluster a large number of people or things into mutually exclusive categories, then an optimum solution in the strictest mathematical sense is not feasible--even with modern high-speed computers. However, the MAXOF Clustering Model yields a near-optimum solution which has passed the test of customer satisfaction.

In using the MAXOF Model, the investigator must make three major decisions. First, he must define a way of expressing the similarity among the things or people to be clustered. The model makes no demands on the form of this overlap function. It can be correlation coefficients, co-variances, cross-training times, distance functions or measures of the homogeneity of regression equations. Any function is legitimate which can be quantified, and which serves the investigator's purpose. Second, the investigator must define an objective function which is to be maximized during the clustering process. For example, the investigator may wish to maximize the average intercorrelation among items within clusters--or to minimize the average squared distance ( $d^2$ ) between items within clusters. Again, there is no restriction on the form of the objective function, except that it be feasible to compute. Third, the investigator must decide on the appropriate number of clusters to report. Problems associated with this decision will be discussed in detail.

The MAXOF Clustering Model takes its name from the concept of MAXimizing an Objective Function, which is its most unique and useful characteristic. The model was first described by Joe H. Ward, Jr., in a paper published in March 1961, under the title "Hierarchical Grouping to Maximize Pay-Off," (Ward, 1961). R. A. Bottenberg and R. E. Christal described in detail a specific application of the model in a paper published this same month entitled "An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency," (Bottenberg & Christal, 1961). Since 1961,

the MAXOF Clustering Model has been applied to many operational problems, with gratifying results. The major purpose of this paper is to describe the model in sufficient detail for readers to determine its applicability to their own problems of interest. For this reason, stress will be given to a discussion of the basic concepts underlying the model, and to a description of previous applications of the model to actual problems. Readers interested in more detail may obtain copies of the references or write directly to one of the authors.

### GENESIS

It all began when Hq USAF asked for development of an improved method for grouping jobs into specialty clusters. Let's expand on this for a moment. The basic management unit in the Air Force is the Air Force Specialty. Every job in the Air Force has been assigned a specialty code number by a local manpower officer. Every man in the Air Force also has been assigned a specialty code number, indicating that he is primarily trained to perform jobs in that specialty.

Enlisted personnel in the Air Force change jobs on an average of once every two years, and can be moved freely from any job to any other job having the same specialty number. When an airman changes jobs, a major cost to the Air Force is the amount of time required for him to reach the same level of proficiency in his new job as he had attained in the job from which he was transferred. If jobs within specialties are not homogeneous, the Air Force pays in two ways. First, it must continually support a large and expensive retraining program; and second, at any point in time, large numbers of men will not have reached proficiency in their current assignment.

It seems clear that jobs should be grouped into specialties in a manner which minimizes the overall cross-training time among jobs within specialties.

Suppose we had the cross-training times among a thousand Air Force jobs. How would we go about clustering them into specialties so as to minimize the average cross-training time among jobs within specialties?

The first urge is to transform the data into a matrix of intercorrelations or  $d^2$ 's, for common clustering techniques usually require one of these data forms as input. But why distort reality? If the goal is to minimize cross-training times among jobs within clusters, then our input matrix should be cross-training time values.

Having settled on the nature of the input matrix, we now turn our attention to the problem of clustering the jobs into specialties so as to meet our objective.

But hold on! As stated, our objective is met before any clustering takes place. If each job is considered to be a separate specialty, then the average cross-training time among jobs within specialties is zero. Furthermore, as the number of clusters (specialties) is reduced, this value must increase to the extent that jobs are not identical.

Yet the whole purpose of clustering jobs in the first place is to provide flexibility to management. The Air Force could not possibly maintain separate training courses for every job. Nor could it move personnel to meet changing priorities unless jobs and individuals are clustered into a limited number of management categories.

It is clear that the larger the number of specialties (management categories), the smaller will be the average cross-training value. At the same time, it also is clear that the smaller the number of specialties, the easier and less expensive it will be to manage the personnel system.

What is needed is an optimum solution for every possible number of clusters; then management can decide on the correct number to implement by weighing retraining costs against the cost of managing a given number of classification categories (specialties).

But how do we obtain an optimum solution for every possible number of clusters? The most direct way would be to have the computer evaluate every possible configuration at every possible level. For example, at the 50 cluster level, the computer would systematically evaluate every possible way of assigning the 1,000 jobs into 50 specialties. It would then report that particular configuration which yields the smallest average cross-training time among jobs within clusters. The same approach would be taken at the 49 cluster level, and so on. Is such an approach feasible? Definitely not. All the computers in the world, working in perfect harmony, could not begin to provide the solution in a lifetime. (See letter in Appendix)!

Another approach must be found--one which approximates an optimum solution, but which is feasible to compute.

It is at this point that the concept of systematically collapsing clusters so as to maximize an objective function comes to mind. The concept is extremely simple, and can be described in terms of its application to the job-grouping problem.

First we must define our objective function, which in this case, is the average cross-training time among jobs within clusters (specialties). We begin with each of the 1,000 jobs in a separate cluster. At this stage the value of the objective function is zero. Next we have the computer evaluate every possible way of reducing the number of clusters from 1,000 to 999. For each of the 499,500 alternatives we can compute the average cross-training time among jobs within clusters.

It turns out that the computer will cluster the two jobs having the smallest average cross-training time. At the next stage we have the computer evaluate every possible way of reducing the number of clusters from 999 to 998, through collapsing two of the existing clusters into a single cluster. It may do this by placing one of the 998 ungrouped jobs in the same specialty as the first pair clustered, so that we end up with a three-job cluster. Or, it may cluster a second pair of similar jobs, so that we end up with two clusters containing two jobs and with each of the remaining 996 clusters containing a single job. All 498,501 possible configurations are evaluated, and that one is selected which yields the smallest value of our objective function. This process of reducing the number of clusters by one at each stage is continued, until all jobs are in a single cluster. In each instance all possible alternatives involving the collapse of two existing clusters are considered, and that alternative is accepted which is evaluated as "best" by the objective function.

Thus, we end up with a solution for each possible number of clusters. We also have exact information concerning the average cross-training time among jobs within specialties at each stage, which can be weighed against management costs in order to arrive at a judgment concerning the optimum number of specialty clusters to maintain.

We were anxious to try this new approach, but unfortunately we did not have a matrix of cross-training times among jobs. However, within a few months Dr Marion E. Hook (Hook & Masser, 1962) had gathered rank-order estimates of the time required for cross-training among 98 existing airman specialties. A complete hierarchical clustering of these data was obtained using the MAXOF model.

Since the 98 specialties had been selected from 40 career fields, we were in a position to compare results of the MAXOF clustering solution at the 40 group stage with the career field membership of these specialties. We found the average cross-training time within groups identified by the MAXOF Model to be markedly smaller than average cross-training time within the official career field groups. We were encouraged with the results, since the clustering technique appeared to operate as predicted.

## IDENTIFYING JOB TYPES

It wasn't long before we found another application of the grouping technique which proved to have considerable pay-off to the Air Force. The Personnel Research Laboratory had been asked to develop improved methods for collecting, analyzing, and reporting information describing enlisted and officer jobs.

We spent the first year studying various job analysis techniques and Air Force needs for job information. The greatest problem was concerned with how to collect information in a form so that it could be quantified and subjected to machine analysis.

Eventually it became clear that a task inventory procedure had greatest potential for satisfying our requirements. Since that decision was made, we have conducted a series of studies concerning how task inventories should be constructed and administered, and how the resulting information should be analyzed and reported.

Our procedures for constructing inventories are relatively straightforward. In the enlisted area, for example, the instrument is simply a list of all the significant tasks performed by individuals working in a single promotion career ladder. That is, it consists of the tasks performed by airmen working at the apprentice, journeyman, supervisor, and superintendent levels in one of the more than 200 career ladders which the Air Force has established for management control. This inventory is administered by test control officers to individuals working in the career ladder at Air Force installations throughout the world. A worker is asked to check those tasks which he performs as part of his normal job, and to indicate how his worktime is distributed across those tasks. He also fills in a background information section, where he indicates such things as his base, command, grade, time on the job, courses taken or equipment worked on.

The completed inventories are sent to the Laboratory, where the data are keypunched and transferred to magnetic tape. Without going into detail, let me simply state that a series of studies has indicated that we get high-quality information using these instruments.

Once the data are in the computer, they are analyzed by fifteen or twenty programs in order to produce reports tailored to meet the needs of various using agencies (Morsh & Christal, In Press).

One program is a general-purpose information retrieval system. It enables the investigator to produce a consolidated description of the work being performed by any specified group of workers. A special group can be identified in terms of values on as many as nine background variables, through use of a series of "and" and "or" statements. For example, one might ask for a description of the work being performed by airmen who have



been in the Air Force less than two years; who bypassed the basic training course; who are less than 19 years of age; who failed to complete high school; and who are working in overseas jobs in the Pacific Air Force Command.

Figure 1 presents the top portion of a typical consolidated job description. This one describes the work being performed by 394 journeymen medical laboratory technicians working in hospitals and clinics throughout the world. Notice the four columns of numbers printed to the right of the task statements. The first column indicates the per cent of members in this group performing the listed task. The second column reports the per cent of worktime spent on the task by individuals who perform it. The third column presents the per cent of the entire groups' worktime spent on the task. This third column is the main element of the description, since it accounts for the worktime of all cases. Tasks in the job description are arranged in descending order based on the magnitude of the values in this column. Thus, "collect blood specimen directly from patients" is the most time-consuming task performed by journeymen laboratory technicians. By the time you read through the third task, you have accounted for 4.30 per cent of the worktime for this group. This is seen by looking at the value in Column 4, which presents the cumulative sum of the values in Column 3.

-----  
Insert Figure 1 about here  
-----

While this job description is an excellent statement of the work performed by journeymen laboratory technicians as a group, it may not be an accurate description of what any one man does. Commanders of local hospitals and clinics can engineer jobs any way they please in order to accomplish their mission effectively. It might be that the jobs in larger hospitals are highly specialized, so that an individual worker performs only a small subset of the tasks. The Air Force wanted to know how work is organized in the field. They wanted to identify and define all of the job types in each career ladder, and find out where each job type exists and who is working in it.

It seemed reasonable that if we had a detailed description of the work performed by each individual in a career ladder, there should be a way to cluster individuals having similar jobs. Hopefully this could be accomplished using the MAXOF Clustering Model.

The first requirement was to develop a matrix of values defining the overlap of each job with every other job. Several measures of job similarity were considered. Two of these potential overlap functions reflected the

extent to which any two jobs contained identical tasks. One was simply the number of tasks in either job, divided into the number of tasks common to both jobs. The second was the per cent of tasks in job A which were also in job B, averaged with the number of tasks in job B which were also in job A. Both of these measures of task overlap were later discarded in favor of a value indicating common worktime.

This common worktime value is obtained for a pair of jobs by summing the smaller of the two time values associated with each task in the inventory. Thus in the example given in Figure 2, the common worktime value is 80 per cent. Notice that by reallocating 20 per cent of the time values on either of the two descriptions, one can perfectly reproduce the other.

-----  
Insert Figure 2 about here  
-----

Once we have computed a matrix of overlap values, we next must define our objective function, or the decision rule to be used in the grouping operation. In this case, we decided to group in a manner that maximizes retention of descriptive accuracy. Thus, we begin with a separate job description for each individual in our sample. At this stage we can perfectly describe the worktime of all workers. Next, we evaluate every possible way of describing the worktime of all workers using N-1 descriptions. At the end of this stage we must describe the jobs of two workers with a single description. To the extent these two jobs are not exactly identical, this single description will make a small error in defining the worktime of the two individuals. It can be shown that this error will be minimized if we select the two jobs having the highest common worktime value. Thus, we can locate the first two jobs to be clustered by identifying the highest value in our original overlap matrix. The composite description for these two jobs is simply an average of the worktime on each task in the inventory.

At the second stage we evaluate every possible way of reducing the number of descriptions by one. The possibilities include locating a third job similar to the first pair, and describing all three with a single description, or finding a second pair of similar jobs to be defined by a composite description. The process is continued, defining the worktime of all cases with one less description at each stage, until we reach the last stage--where we attempt to define all jobs with a single description.

In order to determine the number of job types in the ladder, we normally work backwards through the solution. Ordinarily, we can quickly eliminate the one description stage because of the magnitude of the error. If the error is too large at the two-group stage, we look at the three-group stage. We proceed in this manner until we reach a point where we cannot tell from the error term alone whether the clusters being merged are similar enough to be considered as being in the same job type. In order to help us reach

a decision point, we have the computer provide us with consolidated descriptions of the groups being merged at several stages around the decision point. We may find, for example, that we cannot detect meaningful differences between the two groups being merged at the 25-group stage. However, we may discover that in order to reduce the number of job-type descriptions from 25 to 24, the computer merges two groups which are different in some significant respect. If so, we conclude that there are 25 significant job types, and we have the computer publish consolidated descriptions of the work in each job type.

In the course of obtaining a complete hierarchical grouping of a 2,000-job input matrix, the computer evaluates 4,333,333,000 possible configurations. However, problems of this magnitude are now accomplished on a routine basis without difficulty. Job-type analyses of some forty career ladders have been completed, and in each instance the results have given us a clear picture of the way that work is organized in the field. For example, sixteen clearly defined job types were identified in the medical laboratory career ladder. The reader will find the top tasks from several of these job type descriptions listed in the appendix.

#### CRITERION GROUPING

It wasn't long before we discovered a new application of the MAXOF Clustering Model. In personnel classification and assignment, a primary goal is to predict performance of each individual in a technical training course associated with a particular job area. Even though a battery of tests is routinely administered to Air Force volunteers, it has not been feasible to develop and utilize a separate test composite for predicting the success of each individual in each technical course. Attempts have been made to group related courses into "families" so that a smaller number of predictor composites could be used.

In the Air Force, highly subjective techniques have been used for grouping courses into homogeneous families and for determining the aptitude composite associated with each family. In general, interrelationships among courses have been estimated by intercorrelating the patterns of predictor validities associated with each school. The intercorrelation matrix has been factor analyzed, and schools with similar factor loadings have been grouped. Finally, weights for aptitude composites have been estimated by averaging validity coefficients for the predictor tests against schools in a cluster.

It should be noted that factor analysis is not appropriate for this type of clustering problem, for it is not our goal to explain the common variance in terms of a minimum number of hypothetical constructs. Furthermore, the rank of the matrix of intercorrelations among technical schools (which can only be estimated) does not help us to decide on the optimal number of clusters. Finally, even if the factor-analytic approach did enable us to assign courses into homogeneous groups, we still would need to determine the weights which yield simultaneous, optimum prediction of all criteria within clusters.

After reflecting on the matter, it seemed that the MAXOF Clustering Model might be applicable. Details of the system which was finally worked out would take too much space to describe in this paper. However, they are spelled out in a Technical Documentary Report (Bottenberg & Christal, 1961) which is available upon request. Conceptually, the system begins with k separate least squared regression equations--one for each of k schools. A computing expression has been developed which enables the investigator to determine the overall predictive efficiency obtained using these k separate equations. As the first step, the two schools are clustered whose associated regression equations are most homogeneous, and a single set of least squares weights is developed for simultaneously predicting criterion scores in both schools. Thus, the number of school groups and associated equations is reduced by one. The process of reducing the number of groups and associated equations by one at each step is continued until the one-group stage is reached. In each instance all alternatives are considered, and that alternative is selected which minimizes loss of overall predictive efficiency. The number of groups and equations to be retained is decided by weighing predictive efficiency against the cost of utilizing a given number of prediction composites.

This is a considerably over-simplified description of the actual criterion grouping system. For example, the program enables the investigator to give weight to training costs, personnel quotas, and other factors associated with a particular school. That is, the program may be oriented toward preserving predictive efficiency for those schools where the number of students and the training costs are high relative to other schools.

In those instances where the criterion means and variances are equal, computing expressions for the grouping system are extremely simple. Under this condition, the computer program can easily accomplish a complete hierarchical grouping of nearly a thousand criterion situations, using predictor composites based on as many as a hundred and fifty variables.

We were able to locate criterion information and classification test scores for airmen attending sixty Air Force technical schools. A complete hierarchical grouping of the schools was accomplished. The results revealed

that reduction of the number of predictor equations from 60 to 15 was associated with a drop in overall  $R^2$  from .56 to .50. However, as the number of equations was reduced from 15 to 1, the  $R^2$  dropped from .50 to .38.

#### JUDGMENT ANALYSIS

In 1963, investigators at the Personnel Research Laboratory were experiencing remarkable success in programming a computer to simulate the actions of decision makers (Ward & Davis, 1963). A subject was required to record a series of decisions into the computer via the console typewriter. The subject made each decision after studying relevant information displayed to him by the typewriter. The computer was programmed to capture the policy of the subject in the form of an equation developed with the fixed-X multiple linear regression model. A series of decisions made by the subject was used as the dependent variable, while the independent variables were generated from information provided to the subject on the typewriter. The policy equation was then cross validated against a second series of decisions.

At that time, the concept of policy capturing using the regression model was rather novel. However, more recently we have found policy-capturing to be a powerful approach to many meaningful operational problems. For example, in one study (Christal, 1965) a Hq USAF board of senior officers reviewed descriptions of 3,575 representative officer positions and made decisions concerning the appropriate grade level to be associated with each. In an effort to identify the factors considered by these board members in making their judgments, over a hundred variables were hypothesized and evaluated. Eventually, a nine-predictor equation was developed which accurately reproduced the board's actions. Subsequently, this equation was applied by the computer to determine appropriate grades for an additional 10,000 officers' positions. In other applications the model has been used to develop a mechanized initial assignment system for airmen which duplicates actions previously performed by assignment specialists. A study is planned to use this technique to develop a reassignment model which gives appropriate consideration to job and personnel characteristics. While these applications have been in the military setting, the policy-capturing model might be used to study such diverse properties as the quality of beefstock, the beauty of pictures, the effectiveness of workers or the quality of English compositions. (Christal, 1966)

As one might expect, we have found that individuals sometimes differ quite markedly in their policies concerning a particular type of stimulus. For example, what might be a beautiful picture to one judge may be dull to another. And what might be an excellent composition to one teacher may appear unacceptable to another. The problem, of course, is that all judges do not have the same value system. We find this problem in the military

setting. If one has a board of officers judge the relative acceptability of a series of applicants for the Air Force Academy, for example, there will be considerable disagreement. Some will place high value on previous participation in high school sports and on the physical characteristics of applicants. Others will tend to place more weight on academic aptitudes.

In the past, even though we might find interrater agreement to be low, we have simply averaged across all judges in order to determine final values. However, it should be recognized that even when the level of interrater agreement among an entire sample of judges is low, it might be that the judges could be divided into two or more groups within each of which there is very high agreement. Conventional analysis techniques for determining interrater agreement would fail to detect this situation. It turns out that the criterion grouping application of the MAXOF Clustering Model is ideal for studying similarities and differences in rating policies (Christal, 1963). We begin with a separate equation for each judge, and then we cluster judges with similar policies as measured by the homogeneity of their associated equations. Sometimes we find that judges can be nicely clustered into two or three policy groups. In such an instance, differences in policies are pinpointed for arbitration.

As an illustration, there was a group of ten supervisors in the personnel department of a large government-owned, government-managed research laboratory who had been arguing about promotion standards for six years. Each year they had met for three days to discuss the matter, but without reaching agreement. Dr Robert Stephenson of the U.S. Naval Ordnance Test Station worked with Dr Ward in conducting a study to resolve the problem (Stephenson & Ward, In Press). First they identified 112 items which might be related to promotion potential. Next each supervisor rated the importance of each item for evaluating promotion potential. Following a study of relevant variables, an analysis was performed in which the position of each supervisor was plotted as a point in multi-dimensional space. "Unfortunately," report the authors, "the knowledge of how similar one's position was to somebody else's position did not really help the members of the group to resolve their conflicting opinions. In fact, the analysis tended to focus attention on people relationships (like 'How similar am I to the boss.') rather than policy differences."

Next Stephenson and Ward tried the "JAN" technique (Christal, 1963), which is nothing more than a combination of policy-capturing and the MAXOF Clustering Model. First the investigators described a sample of potential promotees in terms of their score values on relevant factors. Then each of the ten supervisors was asked to rank the entire sample in terms of judged merit. The policy of each supervisor was captured using the multiple-linear regression model. The supervisors were then clustered in terms of the homogeneity of their equations, using the MAXOF Model. Three distinctive policy groups were identified, and three associated joint-policy equations

were developed. These three equations were applied to rank the applicant sample, and the supervisors were asked to discuss and resolve differences in the rank positions of cases resulting from application of these three equations. It is interesting to note that the supervisors spent more time resolving these differences than they did in making their original rankings. However, as a result of this undertaking the supervisors began to understand each other's positions, and found compromise possible. Ultimately, a compromise ranking was arrived at for each controversial case. A single new equation was developed which produced an  $R^2$  of .932. This is almost unbelievable, when one considers that the best single overall equation which could be attained before arbitration produced an  $R^2$  of only .482, and that the best equation for an individual supervisor produced an  $R^2$  of only .848. The authors concluded that the JAN technique was doubly successful. The supervisors gained an understanding of each others positions, and they also reached agreement on a matter over which they had been fighting for six years.

#### BIOLOGICAL TAXONOMIES

As mentioned previously, the MAXOF Clustering Model is ideally suited for establishing biological taxonomies. It yields a completely-nested hierarchical structure based upon optimization of a criterion established by the investigator. The MAXOF Model was used by a New York botanist (unpublished study) to establish a taxonomy of Latin American tapioca plants. The model is now being applied to cluster a group of tropical fish in terms of the similarity of their eating habits. In this instance, the problem turns out to be identical to the job-typing problem. In place of a job description reporting the "per cent worktime" on each of N tasks, we compute a "stomach content" description, reporting the per cent of total stomach content accounted for by each of N foods. Instead of an input matrix of common worktime values, we have an input matrix of common food values. If the volume of food were considered to be a relevant factor, then a matrix of  $d^2$ s could be generated as input which would give weight to differences in volume as well as types of food consumed.

#### INTEREST AREAS AND DOCUMENT GROUPING

The MAXOF Clustering Model was used to group interest areas displayed by scientists at the Personnel Research Laboratory. Results of this study (Tomlinson, 1965) are reported in Figure 3. One of the advantages in obtaining a completely-nested hierarchical solution is revealed in this figure. It is possible to determine a particular sequencing of the items being clustered, such that items appearing in any cluster at any stage are listed next to one another. Thus in Figure 3 the reader can view 67 levels of the hierarchical solution.



After clustering interest areas, this investigator obtained a transpose of the original matrix and clustered the scientists in terms of the similarity of their reading interests. Both solutions were accepted by personnel working in the Laboratory as being a true representation of reality.

In a somewhat related study, an investigator at the Systems Development Corporation in California reported (unpublished study) that the MAXOF Model turned out to be nearly ideal as a basis for a mixed document and word grouping approach to be used in document storage and retrieval.

#### MISCELLANEOUS APPLICATIONS

The MAXOF Clustering Model has been used in several profile analysis studies. For example, it has been used to cluster subjects in terms of their test profiles (Ward & Hook, 1963). The model also has been used to group psychiatric patients in terms of the similarity of their profiles on personal history, socio-economic, and other variables considered relevant to diagnosis and prognosis.

In most studies of this type, some form of a distance function is used as a measure of similarity. There is no problem in applying the MAXOF Clustering Model to group things or people so as to minimize the distances or squared distances among items within clusters. However, distance values are at best rather ambiguous statements, being affected by the number, types, and nature of variables used in their computation. Distance functions should be avoided when more relevant and understandable measures of similarity can be utilized. Nevertheless, there are occasions when more meaningful values cannot be defined. This being the case, the investigator should at least exercise some control over the contribution of variables to the computed distance values. One approach would be to avoid geometric distances altogether, and to substitute measures of perceived distances. A group of experts in the discipline area could be provided with profile descriptions for a subsample of the things or people to be clustered and asked to make direct judgments of the distances among them. The fixed-X multiple linear regression model could then be employed to determine how difference scores on the descriptive variables must be weighed in order to implement the policy of these experts. This equation could be applied to determine the perceived distances among the entire sample of things or people to be grouped. Using this input matrix, clusters could be defined which are likely to have face validity, since they contain items perceived as being close to one another by experts in the discipline area.



## PROGRAM DESCRIPTIONS

The Personnel Research Laboratory has two sets of computer programs available for applying the MAXOF Clustering Model. The most elaborate set contains a large variety of options for computing input matrices, clustering, and generating reports of results. This systems package is designed for execution on an IBM 7040 computer which has a 32K core memory, a 1301 disk file, six addressable tape units, an on-line 1402 reader-punch, an on-line 1403 printer, and an inquiry station. A few of the routines in this package will be described below. For investigators who do not have access to a computer with disk storage, a special set of programs has been written which will accomplish profile, criterion, and job clustering on a smaller scale. Input limitations are a function of core size.

The profile clustering program permits grouping of 1,000 cases. Input is normally from punched cards, and may include up to 928 words of background and history information on each case in addition to the profile data. Profile data may consist of score values on up to 928 variables. Values on a particular profile variable must fit into a field of eight columns, including the sign and decimal point, if required. The system reads and edits input data, assigns case numbers, and writes data on tape. Twelve options are available for computing a matrix of similarities among profiles. These are defined in the appendix. The first will be recognized as the  $d^2$ 's computed from raw scores. The second  $d^2$ 's computed from standardized scores. The standardization routine is part of the basic program and is accomplished automatically if option 2 is selected. Option 3 permits the investigator to introduce weights to be applied to raw score variables. It results in  $d^2$ 's computed from weighted raw scores. Option 4 produces  $d^2$ 's computed from weighted standardized scores. Again, the investigator provides the weights. Options 5 through 8 correspond to options 1 through 4, except that values in the latter matrices are the positive square roots of the values in the former matrices. Thus they might be defined as being  $d$ 's rather than  $d^2$ 's. Options 9 through 12 produce summations of absolute differences of (a) raw scores, (b) standard scores, (c) weighted raw scores, and (d) weighted standard scores, respectively.

Once the selected input matrix has been computed, it may be written on tape or stored on disks ready for immediate grouping. The function of the grouping program is to combine or "collapse" two rows of the matrix at a time until only one final row remains. This collapsing is guided by an "option" or objective function selected by the user. A total of six pre-programmed options are available. The program also provides a way for the user to code and insert his own objective function. Definition of the pre-programmed options are given in the appendix. Ordinarily, profile analysis

is accomplished using the maximizing function associated with options 1 or 2. Thus in option 1, groups are collapsed so as to maximize the average within-group overlap at each stage. In option 2, which we have found best for most purposes except criterion grouping, the program simply collapses the two groups whose members are most similar; that is, for which the average pair-wise between-groups member overlap is highest.

After the grouping has been accomplished, the program enables the investigator to publish several types of tables displaying the results. First, one can obtain a group profile description for any group existing at any stage. The format of a group profile description is illustrated in the appendix. Programs also are available for describing individuals in any group in terms of the history and background data. One can request distributions, means and standard deviations for selected background variables. Many other types of displays are possible which cannot be described due to space limitations. It is suggested that anyone desiring more information about these programs write to one of the authors.

Input to the criterion grouping program is in the form of beta weights and validity coefficients. The appendix includes a note written by Dr Robert A. Bottenberg which demonstrates how an input matrix of "pair-wise loss values" can be grouped with option 1 of the general grouping program in a manner which minimizes over-all loss in predictive efficiency. In the case of equal criterion means and standard deviations, the program can handle nearly a thousand input equations, each involving a common set of not more than 150 predictors. Outputs include (a) the overall  $R^2$  at each stage, (b) the set of raw score regression weights for the new group formed at each stage, and (c) the set of validity coefficients for the new group formed at each stage. Again, several other outputs are available, which can be described to requestors.

The task survey analysis programs are by far the most elaborate, and cannot be described in this paper.

#### SUMMARY

A hierarchical clustering technique has been described which is designed to group people or things into mutually exclusive categories. The input matrix of overlap values may take any form which the investigator selects as representing reality.

The model begins with each of the  $N$  objects in a separate group. The number of groups is reduced by one at each stage, until all objects are in a single group. Choice of the two groups to be collapsed at a given stage is determined by considering all possibilities and selecting that one which

best satisfies an objective function previously established by the investigator. Thus, the model groups objects into every possible number of mutually exclusive clusters, from N to 1. The investigator decides on the appropriate number of clusters to report by considering relevant factors.

Applications of the model described in the paper include: (a) grouping jobs in a manner which minimizes average cross-training time among jobs within clusters; (b) defining a large number of jobs with a fewer number of consolidated job descriptions in a manner which maintains maximum descriptive accuracy; (c) clustering technical schools into families and producing associated prediction equations so as to maintain maximum predictive efficiency; (d) clustering judges in terms of the homogeneity of their policy equations, and producing composite equations for each group accepted; (e) establishing a taxonomy of Latin American tapioca plants; (f) grouping tropical fish in terms of the similarity of their eating habits; (g) grouping reading areas; (h) grouping scientists in terms of the similarity of their reading interests; (i) document grouping; (j) task clustering, and (k) profile analysis.

#### REFERENCES

- Bottenberg, R. A. and Christal, R. E. An iterative technique for clustering criteria which retains optimum predictive efficiency. WADD-TN-61-30. Lackland Air Force Base, Tex.: Personnel Laboratory, Aeromedical Space Division, March 1961.
- Christal, Raymond E. Using the electronic computer to define and implement policy. Proceedings of the 13th Annual Air Force Science and Engineering Symposium. Tullahoma, Tennessee, September 1966.
- Christal, Raymond E. Officer grade requirements I. Overview. PRL-TDR-65-15. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, September 1965
- Christal, Raymond E. JAN: A technique for analyzing group judgment. PRL-TDR-63-3. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, February 1963.
- Hook, Marion E. and Massar, Richard S. Rankorder estimates of the time required for cross-training among 98 airmen specialties. PRL-TDR-62-15. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, August 1962.
- Morsh, Joseph E. and Christal, Raymond E. Impact of the computer on job analysis in the United States Air Force. PRL-TR-66- (IN PRESS). Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division.
- Stephenson, Robert W. and Ward, Joe H., Jr. Applied use of judgment analysis (JAN) to help a policy group to resolve a controversial issue. Unpublished manuscript, available from authors.
- Tomlinson, Helen. Classification of information topics by clustering interest profiles. PRL-TR-65-19. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, November 1965.
- Tomlinson, Helen. Defining technical information needs for a research laboratory. PRL-TR-65-4. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, March 1965.
- Ward, Joe H. Jr., and Davis, Kathleen. Teaching a digital computer to assist in making decisions. PRL-TR-63-16. Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, June 1963.

- Ward, J. H., Jr., and Hook, Marion E. Application of a hierarchical grouping procedure to a problem of grouping profiles. Educ. Psychol. Measmt. (1963, 23, 69-6).
- Ward, J. H., Jr. Hierarchical grouping to maximize pay-off. WADD-TN-61-29 Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, March 1961.

## APPENDIX

Contents of Appendix are as follows, listed in order of appearance:

1. Requirements for computer evaluation of objective function on pooling a large number of objects.
2. Definition of input matrices for profile grouping.
3. Format for description of group profile.
4. Definition of grouping process and collapsing formulas.
5. Conditions for use of options 4 and 6 for grouping in terms of square multiple correlation coefficients.
6. Listing of top tasks from six medical Laboratory Technician job type descriptions.

DEPARTMENT OF THE AIR FORCE  
HEADQUARTERS 6570TH PERSONNEL RESEARCH LABORATORY (AFSC)  
LACKLAND AIR FORCE BASE, TEXAS 78236



REPLY TO  
ATTN OF

PRMM *gim*

26 September 1966

SUBJECT

Requirements for Computer Evaluation of Objective Function, on Pooling a Large Number of Objects

TO PRB (Dr. Christal)

1. The most general statement of the problem is to determine an estimate of machine time required to perform the evaluations for all possible groupings of 1,000 objects. The enumeration of all such groupings appears to be a difficult problem. Therefore, what follows is limited to the enumeration of a subset of these groupings. Any time estimate made on the basis of evaluating the objective function for all groupings in this subset will be a gross underestimate of the time requirement for grouping in all possible ways. The subset in question contains any grouping such that: a. there are 500 groups, and b. there are exactly two objects in each group. Define this subset as S.

2. To enumerate the groupings of 1,000 objects into 500 partitions of two objects each, first consider the simple problem of enumerating the groupings of 4 objects into two partitions of two objects each. There are three such groupings, (1,2:3,4), (1,3:2,4), and (1,4:2,3). Note that the lead element in the first partition is the i.d. number 1., and we can make this true for any arbitrary arrangement of i.d. numbers into partitions by reordering the object i.d. numbers within a partition and the order of the partitions without in any way altering the unique grouping. Next consider grouping 6 objects into three partitions of two objects each. Again let the i.d. number 1 be the lead element of the first partition. There are five other i.d. numbers which can be used to fill out the first partition. Then, as in the case of grouping four objects into two partitions of two each, there are three ways of partitioning the remaining four objects for each of the five ways of completing the first partition. So there are  $5 \times 3 = 15$  ways of grouping 6 objects into three partitions of two objects each. The same general argument holds for the problem of putting 8 objects into four partitions of two each. There are 7 i.d. numbers which can be used to complete the first partition after assigning i.d. number 1 to the lead element of the first partition. For each of the 7 ways of completing partition 1, there are  $5 \times 3 = 15$  ways of assigning the remaining 6 objects to three partitions of two objects each. It can be seen by induction that for grouping 1,000 objects into 500 partitions of two objects each, there are  $999 \times 997 \times \dots \times 3 \times 1$  different ways,  $= (1000! / (500! \times 2^{500})) = \text{approximately } 10^{1289}$ .

3. There are approximately  $3 \times 10^7$  seconds per year. Assuming that the cycle time for a computer is about  $10^{-9}$ , and that one evaluation of the objective

function could be performed on each machine cycle, then a computer operating for one year could compute  $3 \times 10^{16}$  evaluations. If the total number of evaluations were split up so that separate computers could perform different subsets of the evaluations, it would require  $10^{1289}/3 \times 10^{16}$  - approximately  $10^{1272}$  computers running continuously for one year to evaluate all the objective functions for the subset S.

*Robert A. Bottenberg*

ROBERT A. BOTTENBERG  
Chief, Mathematical and Statistical  
Analysis Branch



## DEFINITION OF INPUT MATRICES FOR PROFILE GROUPING

A. Input. (1) A tape file or deck of cards containing scores on a set  $p$  variables for each of  $n$  cases.

(2) A set of  $p$  weights, one for each of the  $p$  variables.

When weights are unspecified, they are assumed to be equal to 1.00.

B. Functions of the Program. This program computes a matrix of profile similarities, ready for input in the PRL grouping programs. The input matrix is symmetric and contains  $n^2$  values. The elements of the matrix are computed according to one of the twelve computing expressions listed below, depending upon the option selected.

### C. Definition of Terms.

Let  $X_{ij}$  = score on variable  $j$  for person  $i$ , or the  $j^{\text{th}}$  element in the  $i^{\text{th}}$  record.  $j \leq 1000, i \leq 1000$

$n$  = the upper value for  $i$ .

$p$  = the upper value for  $j$ .

$A_k$  = weight to be applied to variable  $k$ .

$\sigma_k$  = standard deviation of variable  $k$ .

$$= \sqrt{\frac{\sum_{i=1}^n X_{ik}^2}{n} - \left( \frac{\sum_{i=1}^n X_{ik}}{n} \right)^2}$$

D. Computing expressions. Twelve options are programmed

as follows:

$$1. DU_{ij}^2 = \sum_{k=1}^p (X_{ik} - X_{jk})^2$$

$$= d^2$$

$$2. DUS_{ij}^2 = \sum_{k=1}^p \left( \frac{X_{ik} - X_{jk}}{\sigma_k} \right)^2$$

$$= d^2 \text{ using standardized scores}$$

$$3. DW_{ij}^2 = \sum_{k=1}^p \left[ A_x (X_{ik} - X_{jk})^2 \right]$$

$$= d^2 \text{ computed from weighted raw scores}$$

$$4. DWS_{ij}^2 = \sum_{k=1}^p \left[ A_k \left( \frac{X_{ik} - X_{jk}}{\sigma_k} \right)^2 \right]$$

$$= d^2 \text{ computed from weighted standard scores}$$

$$5. DU_{ij} = \sqrt{DU_{ij}^2}$$

$$6. DUS_{ij} = \sqrt{DUS_{ij}^2}$$

$$7. DW_{ij} = \sqrt{DW_{ij}^2}$$

$$8. DWS_{ij} = \sqrt{DWS_{ij}^2}$$

$$9. ADU_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}|$$

$$= \text{Summation of absolute differences in raw scores}$$

$$10. \quad ADUS_{ij} = \sum_{k=1}^p \left| \frac{X_{ik} - X_{jk}}{\sigma_k} \right|$$

= Summation of absolute differences of standard scores

$$11. \quad ADW_{ij} = \sum_{k=1}^p \left[ A_k \left( |X_{ik} - X_{jk}| \right) \right]$$

= Summation of weighted absolute differences of raw scores

$$12. \quad ADWS_{ij} = \sum_{k=1}^p \left[ A_k \left( \left| \frac{X_{ik} - X_{jk}}{\sigma_k} \right| \right) \right]$$

= Summation of weighted absolute differences in standardized scores.

NOTE: The general program is written so that if options 1 and 5 are required simultaneously, the program will obtain both matrices at one time; similarly for options 2 and 6, 3 and 7, and 4 and 8.

# FORMAT FOR DESCRIPTION OF GROUP PROFILE

HEADER INFORMATION -----  
 -----

PROFILE DESCRIPTION, CASES = -, VARIABLES = -, MERS = -.  
 KPATH ORDER FROM TO , GROUP STAGE  
 OVERLAP OPTION XXXXX , GROUPING OPTION XXXXX

		RATIO OF SAMPLE VARIANCE TO GROUP VARIANCE (A/B) - - - - -			
		SAMPLE VARIANCE (A) - - - - -			
		GROUP VARIANCE (B) - - - - -			
		SAMPLE MEAN - GROUP MEAN - - - - -			
		GROUP MEAN - - - - -			
VAR NO.	VARIABLE TITLE				
X	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXX	-XXXXX	XXXXX	XXXXX
X	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXX	-XXXXX	XXXXX	XXXXX
X	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXX	XXXXX

## DEFINITION OF GROUPING PROGRAM AND COLLAPSING FORMULAS

**FUNCTION:** To combine rows of matrix (stored on disk), two at a time, until only one final value remains. This process is called "collapsing the matrix." Two methods are available for selection of the sequence of rows to be combined:

1. **MAXIMIZING Process:** The largest value  $V_{ij}$  in the matrix is searched for each time, and when found, the indices of its position in the matrix ( $i$  and  $j$ ) become the two rows to be combined. Thus, if the numerically largest matrix element is in 123<sup>rd</sup> cell of row 45, then row 123 will be combined with row 45.
2. **MINIMIZING Process:** Similar to the maximizing process except the smallest value is searched for each time. Once either the minimizing or maximizing process is selected (via control card) it remains in effect for all collapses of the entire matrix.

The value selected to determine each collapse is called the BEST value for that collapse; each collapse is called a STAGE. The two row numbers are called the IBEST and JBEST indices for that STAGE. The rows for each STAGE are combined together according to a pre-determined formula. It can be shown that after a combination, no value can be generated which is greater than BEST for that stage if maximizing, or smaller than BEST if minimizing. The row indicated by the larger index is always collapsed into the row with the smaller index. Hence, if BEST is found at 123 and 45, then the new values generated will be restored into row 45, and row 123 will be considered deleted from the matrix. If the matrix was  $m \times m$  to start, then the first collapse is called STAGE  $(m - 1)$ , the next is called

STAGE (  $m - 2$ ), down to stage 1. The original matrix is destroyed in the process.

The original matrix normally was created by one of the OVERLAP programs. Its values really form a "triangular" matrix but it was found that by reflecting the values on the other side of the diagonal, that the grouping program could be made to operate very rapidly. This is done by a "delayed updating" process developed in contract AF 41-(609)-1982. Details will not be repeated here. Actually, a table of BEST values is maintained in core so as to avoid searching the entire matrix on disk for each collapse. The BEST table, the generated weights (number of rows combined into a given row) and order of collapses is also maintained by the program and used as part of the technique.

#### COLLAPSING FORMULAS:

The user must choose one of these and punch its identification number on a control card. The entire collapse of the total matrix then occurs according to the chosen option. In all the below:

$i$  = Lower numbered row of any pair of rows being combined.

New values are restored into row  $i$ .

$j$  = Higher numbered row of any pair of rows being combined.

Row  $j$  is then deleted from matrix.

$k$  = Successive values in a row, where  $K = 1, 2, \dots, m$  and  $m$  is order of the matrix, except no value is computed for the diagonal element  $k = i$ .

$V_{ki}$  = Old values obtained from row  $i$ .

$V_{kj}$  = Old values obtained from row  $j$ .

$V'_{ki}$  = New value due to the combination and restored into position  $k$  or row  $i$ .

$W_x$  = Weight of a row, where  $x$  can be  $i, j$ , or  $k$  depending on the option selected.

At the final stage of collapse, the weight value will be equal to the sum of weight row of the matrix.

#### Collapse Option 1

$$V'_{ki} = \frac{V_{ki} (W_i + W_k) + V_{kj} (W_j + W_k) - V_{ij} W_k}{W_i + W_j + W_k}$$

#### Collapse Option 2

$$V'_{ki} = \frac{V_{ki} W_i + V_{kj} W_j}{W_i + W_j}$$

#### Collapse Option 3 (when MAXIMIZING)

$$V'_{ki} = \text{larger of } V_{ki} \text{ and } V_{kj}$$

#### Collapse Option 3 (when MINIMIZING)

$$V'_{ki} = \text{smaller of } V_{ki} \text{ and } V_{kj}$$

#### Collapse Option 4

$$V'_{ki} = \text{A value generated by user written code incorporated in the GROUP program.}$$

In all options, the new weight of row  $i$  (designated  $W_i$ ) will be the sum of the old  $W_i$  plus  $W_j$ .

### SPECIAL CALCULATIONS

#### TASK INVENTORY and/or PROFILE ANALYSIS

AVERAGE WITHIN =  $V'_{ii}$

$$V'_{ii} = \frac{V_{ii} W_i^2 + V_{jj} W_j^2 + 2V_{ij} W_i W_j}{(W_i + W_j)^2}$$

WHERE:  $V_{ii}$  was the previous "average within" for all rows collapsed into row i (usually, starts at 100%).

$V_{jj}$  is like  $V_{ii}$  except for row j.

$V_{ij}$  is BEST, the value used as the criteria to select these two rows for combination.

$W_i$  and  $W_j$  are the weights of the respective rows before combining.

### REGRESSION EQUATION ANALYSIS

When collapse Option 1 is utilized, a value  $R_k^2$  will be computed. For the initial collapse  $R_g^2$  will be computed in OVLAP1 by the formula:

$$R_g^2 = \frac{\sum_{j=1}^g W_j r_{jj}}{\sum_{j=1}^g W_j}$$

Where  $r_{jj}$  was a calculated value of the squared multiple correlation coefficient.

For every collapse thereafter,  $R_k^2$  will be computed by the formula:

$$R_k^2 = R_{k+1}^2 - V_{ij}$$



Where  $R_k^2$  is the overall  $R^2$  at stage K and  $V_{ij}$  is the matrix element located by the i and j indices.

The value  $R_k^2$  is the output in the place of AVERAGE within.

CONDITIONS FOR USE OF OPTIONS 4 AND 6  
FOR GROUPING IN TERMS OF SQUARED  
MULTIPLE CORRELATION COEFFICIENTS

1. General. Assume an input matrix,  $V$ , exists. Both options 4 and 6 will use  $V$ , seek for the smallest value in  $V$ , and then update. The matrix  $V$  is assumed to be symmetric. If the minimum value in  $V$  is the element  $v_{ij}$  where  $i$  is less than  $j$ , then row and column  $i$  are updated and the existing row and column  $j$  will be disregarded in subsequent operations. After the elements in row and column  $i$  have been updated, the weight  $w_i$  is updated and the existing weight  $w_j$  is subsequently disregarded. The expression for updating element  $k$  in column  $i$  depends on the option. For option 4,

$$v'_{ki} = (v_{ik}w_i + v_{jk}w_j)/(w_i + w_j),$$
 where  $i, j$  identify the position of the smallest element,  $v_{ij}$  in  $V$ ;

for option 6,

$$v'_{ki} = (v_{ik}(w_i + w_k) + v_{jk}(w_j + w_k) - v_{ij}w_k)/(w_i + w_j + w_k),$$
 where  $i, j$  identifies the minimum in matrix  $V$ .

For both options, the updated  $w_i$  is given by  $w'_i = w_i + w_j$ .

2. Assumptions.

a. Proportionality of sums of squares and cross-products of predictor matrices between the initial groups. Equality of predictor intercorrelation matrixes for initial groups is necessary but not sufficient, since the solution for a set of beta weights in combined groups would involve the predictor correlation matrix for the combined group, and even though these matrices are equal for the separate groups, the combined group predictor correlation matrix will not in general be a weighted sum of the separate matrices unless the sums of squares and cross-products matrices for separate groups are proportional. Proportionality also implies that the predictor mean for a given variable is constant from group to group, similarly for the predictor s.d.

b. Equality of criterion variable means across initial groups.

c. Equality of criterion variable s.d. across initial groups.

### 3. Definitions.

- a.  $B_i$ , a row vector of beta weights (standard partial weights) in which the (p)th element is the weight for predictor p in initial group i.
- b.  $B$ , a matrix in which the rows are the  $B_i$ .
- c.  $T_j$ , a column vector of validity coefficients in which the (p)th element is the validity of predictor p for the criterion in initial group j.
- d.  $T$ , a matrix in which the columns are the  $T_j$ .
- e.  $R$ , a square matrix,  $BT$ .  $R$  is theoretically symmetrical but will, in general, fail to be symmetrical due to inaccuracy in solving for the  $B_i$ .
- f.  $w_i$ , weights. Initial values of the  $w_i$  are set at the corresponding  $N_i$  (number of criterion observations) for option 6 with unequal  $N_i$ , and set at 1 when option 6 is used for an equal  $N$  case and for option 4.
- g.  $w$ , the sum over i of initial values of  $w_i$ .
- h.  $V$ , a symmetric matrix in which the element  $v_{ij}$  is obtained from elements  $r_{ii}$ ,  $r_{jj}$ ,  $r_{ij}$ , and  $r_{ji}$  of matrix  $R$ .  
$$v_{ij} = (1/w)(w_i w_j (r_{ii} + r_{jj} - r_{ij} - r_{ji})) / (w_i + w_j),$$
 where the values of  $w_i$  and  $w_j$  are the initial values.

4. Proof that option 6 combines groups so as to minimize loss in over-all predictive efficiency, given the input matrix  $V$  and the updating procedure described in paragraph 1.  
Method: (1) Assume that grouping has occurred and that at the end of this stage the element  $v_{ij}$  is found to be best (minimum) in the updated matrix; (2) That  $v_{ij}$ ,  $v_{ik}$ , and  $v_{jk}$  are the over-all loss in predictive efficiency when the i,j cluster is combined, when the i,k cluster is combined, and when the j,k cluster is combined respectively; (3) Then to show that  $v'_{ki}$  as given by the updating expression will be the over-all loss in predictive efficiency when the i,j cluster is combined with cluster k; and (4) To show that the initial values in the  $V$  matrix represent the over-all loss in predictive efficiency when two of the initial groups are combined. If (3) and (4)

are demonstrated, then by induction the elements in matrix  $V$  at the end of any grouping stage will be the loss in predictive efficiency, over all, when the two clusters identified by the subscripts are combined.

a. Denote the squared multiple for clusters  $i$ ,  $j$ , and  $k$ ; cluster  $i,j$ ; cluster  $i,k$ ; cluster  $j,k$ ; and cluster  $i,j,k$  as  $R_i^2$ ,  $R_j^2$ ,  $R_k^2$ ,  $R_{i,j}^2$ ,  $R_{i,k}^2$ ,  $R_{j,k}^2$ , and  $R_{i,j,k}^2$ . Denote the updated values of the weights for clusters  $i$ ,  $j$ , and  $k$  after  $s$  stages as  $w_i$ ,  $w_j$ ,  $w_k$ , and  $w_r, \dots, w_g, w_b$ , etc. as the weights associated with the initial group indicated by the subscript.

b. The over-all predictive efficiency after  $s$  stages is  $(1/w)(w_i R_i^2 + w_j R_j^2 + w_k R_k^2 + C)$ , where  $C$  is the weighted sum of squared multiples for other clusters.

c. The over-all predictive efficiency, if at the next stage clusters  $i$  and  $j$  were combined, would be  $(1/w)((w_i + w_j)R_{i,j}^2 + w_k R_k^2 + C)$ .

d. The corresponding loss in over-all predictive efficiency is  $(1/w)(w_i R_i^2 + w_j R_j^2 - (w_i + w_j)R_{i,j}^2)$ .

e. By analogy, the loss in over-all predictive efficiency, if  $i$  and  $k$  are combined after stage  $s$ , is  $(1/w)(w_i R_i^2 + w_k R_k^2 - (w_i + w_k)R_{i,k}^2)$ .

f. By analogy, the loss in over-all predictive efficiency, if  $j$  and  $k$  were combined after stage  $s$ , is  $(1/w)(w_j R_j^2 + w_k R_k^2 - (w_j + w_k)R_{j,k}^2)$ .

g. By analogy, the loss in over-all predictive efficiency, if at stage  $s+2$  cluster  $k$  is combined with the  $i,j$  cluster which is assumed to have been combined at stage  $s+1$ , is  $(1/w)((w_i + w_j)R_{i,j}^2 + w_k R_k^2 - (w_i + w_j + w_k)R_{i,j,k}^2)$ .

h. Now assume that element  $v_{ij}$  contains the quantity shown in step d after  $s$  stages,  $v_{ik}$  contains the quantity in step e,  $v_{jk}$  contains the quantity shown in step f, and that  $v_{ij}$  is the minimum in the updated  $V$  matrix. Then show that the updating expression  $v'_{ki} = (1/(w_i + w_j + w_k))(v_{ik}(w_i + w_k) + v_{jk}(w_j + w_k) - v_{ij}w_k)$  will yield the value of the quantity shown in step g.

i. Substituting their assumed values for  $v_{ik}$ ,  $v_{jk}$ , and  $v_{ij}$ ,  

$$v'_{ki} = (1/(w(w_1 + w_j + w_k)))(w_1^2 R_1^2 + w_j^2 R_j^2 + w_k(w_1 + w_j + 2w_k)R_k^2 \\ - (w_1 + w_k)^2 R_{1,k}^2 - (w_j + w_k)^2 R_{j,k}^2 + w_k(w_1 + w_j)R_{1,j}^2).$$

j. Denote  ${}_s B_i$  the vector of beta weights for cluster i after s stages,  ${}_s B_j$  and  ${}_s B_k$  similarly;  ${}_s T_i$ ,  ${}_s T_j$ ,  ${}_s T_k$  are vectors of validity coefficients after stage s for clusters i, j, and k respectively.

k. Let

${}_{s+1} B_{i,j}$  be the vector of beta weights for the combined i,j cluster, if clusters i,j were combined on stage s+1, and by proportionality assumption =  $(1/(w_1 + w_j))(w_1 \cdot {}_s B_i + w_j \cdot {}_s B_j)$ ;

${}_{s+1} B_{i,k}$  be the vector of beta weights for the combined i,k cluster, if on stage s+1 clusters i and k are combined, =  $(1/(w_1 + w_k))(w_1 \cdot {}_s B_i + w_k \cdot {}_s B_k)$ ;

${}_{s+1} B_{j,k}$  be the vector of beta weights for the combined j,k cluster, if on stage s+1 clusters j and k are combined, =  $(1/(w_j + w_k))(w_j \cdot {}_s B_j + w_k \cdot {}_s B_k)$ ;

${}_{s+2} B_{i,j,k}$  be the vector of beta weights for the combined i,j,k cluster, if on stage s+1 clusters i and j are combined and on stage s+2 cluster k is combined with the i,j cluster formed on stage s+1, =  $(1/(w_1 + w_j + w_k))(w_1 \cdot {}_s B_i + w_j \cdot {}_s B_j + w_k \cdot {}_s B_k)$ ;

and define vectors of validity coefficients similarly.

l. Then,

$$R_i^2 = {}_s B_i \cdot {}_s T_i;$$

$$R_j^2 = {}_s B_j \cdot {}_s T_j;$$

$$R_k^2 = {}_s B_k \cdot {}_s T_k;$$

$$R_{i,j}^2 = {}_{s+1} B_{i,j} \cdot {}_{s+1} T_{i,j}, \text{ and substituting the s stage vectors for the s+1 stage vectors as in step k, } = (1/(w_1 + w_j)^2)(w_1^2 \cdot {}_s B_i \cdot {}_s T_i + w_j^2 \cdot {}_s B_j \cdot {}_s T_j + w_1 w_j \cdot {}_s B_i \cdot {}_s T_j + w_1 w_j \cdot {}_s B_j \cdot {}_s T_i);$$

$$R_{1,k}^2 = s+1B_{1,k} \cdot s+1T_{1,k} \text{ or } = (1/(w_1 + w_k)^2)(w_1^2 \cdot sB_1 \cdot sT_1 + w_k^2 \cdot sB_k \cdot sT_k + w_1w_k \cdot sB_1 \cdot sT_k + w_1w_k \cdot sB_k \cdot sT_1);$$

$$R_{j,k}^2 = s+1B_{j,k} \cdot s+1T_{j,k} \text{ or } = (1/(w_j + w_k)^2)(w_j^2 \cdot sB_j \cdot sT_j + w_k^2 \cdot sB_k \cdot sT_k + w_jw_k \cdot sB_j \cdot sT_k + w_jw_k \cdot sB_k \cdot sT_j); \text{ and}$$

$$R_{1,j,k}^2 = s+2B_{1,j,k} \cdot s+2T_{1,j,k} \text{ or } = (1/(w_1 + w_j + w_k)^2)(w_1^2 \cdot sB_1 \cdot sT_1 + w_j^2 \cdot sB_j \cdot sT_j + w_k^2 \cdot sB_k \cdot sT_k + w_1w_j \cdot sB_1 \cdot sT_j + w_1w_j \cdot sB_j \cdot sT_1 + w_1w_k \cdot sB_1 \cdot sT_k + w_1w_k \cdot sB_k \cdot sT_1 + w_jw_k \cdot sB_j \cdot sT_k + w_jw_k \cdot sB_k \cdot sT_j).$$

m. Substituting from step l the expressions for  $R_{1,j}^2$ ,  $R_{j,k}^2$ ,  $R_{1,k}^2$ ,  $R_{1,j,k}^2$  into the expression for  $v'_{ki}$  in step i,  $v'_{ki} = (1/(w(w_1 + w_j + w_k)))(w_k w_1^2 \cdot sB_1 \cdot sT_1 / (w_1 + w_j) + w_k w_j^2 \cdot sB_j \cdot sT_j / (w_1 + w_j) + w_k(w_1 + w_j) \cdot sB_k \cdot sT_k + w_1w_jw_k \cdot sB_1 \cdot sT_j / (w_1 + w_j) + w_1w_jw_k \cdot sB_j \cdot sT_1 / (w_1 + w_j) - w_1w_k \cdot sB_1 \cdot sT_k - w_1w_k \cdot sB_k \cdot sT_1 - w_jw_k \cdot sB_j \cdot sT_k - w_jw_k \cdot sB_k \cdot sT_j).$

n. Substituting from step l the expressions for  $R_{1,j}^2$ ,  $R_{j,k}^2$ , and  $R_{1,j,k}^2$  into the expression for loss in over-all predictive efficiency in step g and evaluating, the quantity is identical to the value of  $v'_{ki}$  derived in step m.

o. Therefore, if it is assumed that the updated V matrix after s stages contains elements which are the loss in over-all predictive efficiency which would be required at stage s+1 if the two clusters indicated by the row and column subscripts of an element were combined, then the expression for updating the elements in the column and row in which the minimum is found will, in fact, give the loss in over-all predictive efficiency when some other cluster, k, is combined on stage s+2 with the combined i, j cluster formed at stage s+1.

p. To prove that the input matrix V contains elements which are the loss in over-all predictive efficiency when a pair of initial groups are combined, let  $A_{i,j}^2$  be the squared multiple for the two group cluster consisting of initial groups i and j. The over-all predictive efficiency for the full interaction model is  $(1/w)(w_i r_{ii} + w_j r_{jj} + K)$ , where K

is the weighted sum of other elements on the diagonal of matrix R. The over-all predictive efficiency, when i and j are combined is

$$(1/w)(w_i + w_j)A_{i,j}^2 + K);$$

and the loss in over-all predictive efficiency is

$$(1/w)(w_i r_{ii} + w_j r_{jj} - (w_i + w_j)A_{i,j}^2).$$

The vector of beta weights for the i,j cluster is, because of proportionality,

$$(1/(w_i + w_j))(w_i B_i + w_j B_j);$$

and the validity vector for the i,j cluster is

$$(1/(w_i + w_j))(w_i T_i + w_j T_j);$$

hence,

$$\begin{aligned} A_{i,j}^2 &= (1/(w_i + w_j)^2)(w_i^2 B_i T_i + w_j^2 B_j T_j + w_i w_j B_i T_j + w_i w_j B_j T_i) \\ \text{or} &= (1/(w_i + w_j)^2)(w_i^2 r_{ii} + w_j^2 r_{jj} + w_i w_j r_{ij} + w_i w_j r_{ji}). \end{aligned}$$

Substituting this value in the expression for the loss in over-all predictive efficiency,

$$\text{loss} = (w_i w_j / (w(w_i + w_j)))(r_{ii} + r_{jj} - r_{ij} - r_{ji}),$$

which is also the expression for the element  $v_{ij}$  in the input matrix (see paragraph 3, item h).

q. Note: If all the initial group n's are equal, the initial values of  $w_i$  could either all be set equal to this common value or the  $w_i$  all set equal to 1 initially. In the latter case, the  $w_i$ , etc., would be just the number of initial groups currently contained in cluster i and w would be the number of initial groups. The expressions for the vectors of beta weights and validity coefficients would still be correct.

5. Proof that option 4 groups so as to minimize the average pairwise loss in predictive efficiency, given an input matrix V and the expression for updating elements in the row and column in which the minimum of the updated V matrix is located,  $v'_{ki} = (v_{ik}w_i + v_{jk}w_j)/(w_i + w_j)$ , where  $v_{ij}$  is the minimum in V. Definition: A pairwise loss in predictive efficiency is the reduction in  $R^2$  when an interaction model for only two of the initial groups is reduced to a common model for the two groups.

If at the end of  $s$  stages there are  $W_i$  of the initial groups in cluster  $i$  and  $W_k$  of the initial groups in cluster  $k$ , then the average loss being considered is the average of  $W_i \cdot W_k$  losses, where each loss reflects the combining of one member of cluster  $i$  with one member of cluster  $k$ . Method: (1) Assume that at the end of  $s$  stages the  $V$  matrix has been updated so that  $v_{ij}$  contains  $2/w$  times the average pairwise loss which will occur on stage  $s+1$  if clusters  $i$  and  $j$  are combined,  $v_{ik}$  contains  $2/w$  times the average pairwise loss which will occur on stage  $s+1$  if clusters  $i$  and  $k$  are combined, and  $v_{jk}$  contains  $2/w$  times the average pairwise loss which will occur on stage  $s+1$  if clusters  $j$  and  $k$  are combined; (2) Assume that  $v_{ij}$  is the minimum in the updated  $V$ ; (3) Then show that  $v'_{ki}$  will be  $2/w$  times the average pairwise loss which will occur on stage  $s+2$  if cluster  $k$  is combined with the new cluster called  $i$  which was formed at stage  $s+1$ ; (4) Then show that an element of the input matrix is  $2/w$  times the loss for a pair of initial groups; (5) So, by induction, all elements of the updated  $V$  matrix will contain  $2/w$  times the average pairwise loss when all possible pairs are formed by putting members of the cluster identified by the row subscript with members of the cluster identified by the column subscript of the element of the updated  $V$  matrix.

a. By assumption (1) under "Method," there are  $W_i \cdot W_k$  losses involved in  $v_{ik}$ ; hence,  $W_i \cdot W_k (w/2) v_{ik}$  is the sum of  $W_i \cdot W_k$  losses when members of cluster  $i$  are paired with members of cluster  $k$ .

b. Similarly,  $W_j \cdot W_k (w/2) v_{jk}$  is the sum of  $W_j \cdot W_k$  losses when members of cluster  $j$  are paired with members of cluster  $k$ .

c. The total number of losses to be considered when members of cluster  $i$  are paired with members of cluster  $k$  and members of cluster  $j$  are paired with members of cluster  $k$  is  $W_i \cdot W_k + W_j \cdot W_k = W_k (W_i + W_j)$ .

d. Therefore,  $(W_i \cdot W_k (w/2) v_{ik} + W_j \cdot W_k (w/2) v_{jk}) / (W_k (W_i + W_j))$   
 $= (w/2) (v_{ik} W_i + v_{jk} W_j) / (W_i + W_j)$  or  $= (w/2) v'_{ki}$  is the average pairwise loss when members of cluster  $i$  are paired with members of cluster  $k$  and members of cluster  $j$  are paired with members of cluster  $k$ .



e. Hence,  $v'_{ki} = (2/w)$  times the average loss which would occur on stage  $s+2$  if clusters  $i$  and  $j$  existing at the end of stage  $s$  are combined on stage  $s+1$ .

f. The two-group interaction model squared multiple for initial groups  $i$  and  $j$  with equal  $n$ 's is  $(1/2)(r_{ii} + r_{jj})$ .

g. The vector of beta weights for the combination of initial groups  $i$  and  $j$  is  $(1/2)(B_i + B_j)$ ; the vector of validity coefficients is  $(1/2)(T_i + T_j)$ ; the squared multiple is then

$$(1/4)(B_i T_i + B_j T_j + B_i T_j + B_j T_i) = (1/4)(r_{ii} + r_{jj} + r_{ij} + r_{ji}).$$

h. The loss for pairing groups  $i$  and  $j$  is

$$(1/4)(r_{ii} + r_{jj} - r_{ij} - r_{ji}) = (w/2)v_{ij}$$

(see paragraph 3h).

i. Hence, the element  $v_{ij}$  of the input matrix  $V$  is  $(2/w)$  times the pairwise loss for initial groups  $i$  and  $j$ .

j. Since steps a through e do not use the assumption of proportional sums of squares and cross-product predictor matrices and of common criterion means and s.d.'s, option 4 can be shown to minimize average loss in pairwise  $R^2$  if some other input matrix  $V$  can be constructed which reflects loss in pairwise  $R^2$  with an expression using the elements of matrix  $R$  and other information so as not to involve these assumptions. However, initial values  $w_i$  must be set at 1 even if another input matrix  $V$  is used.

Top Tasks from Six Medical Laboratory  
Technician Job Types

Bio Chemistry Job Type

Perform Liver Function Tssts  
Perform NPN and BUN Tests  
Operate Spectro Photometer  
Perform Calcium and Phosphorus Tests  
Perform Total Protein and A G Ratio  
Total Cholesterol and Esters Test  
Utilize Colormetric Procedure  
Perform URIC Acid Tests  
Perform Carbon Dioxide Determinations  
Perform Enzyme Analyses  
Perform Chlorides Tests  
Prepare Reagents and Standards  
Perform Electrolyte Determinations  
Collect Blood Specimens Directly from Patients  
Perform Carbohydrates Tolerance Tests  
Operate Flame Photometer  
Perform Creatinine Tests  
Prepare Reagents  
Perform Prothrombin Time Test  
Prepare Solutions and Standards  
Clean Area Equipment Aseptically  
Separate Serum from Blood  
Prepare and Process Specimens  
Centrifuge and Separate Serum from Clot  
Utilize Titrimetric Procedure

Blood Bank Job Type

Crossmatch Blood  
Test Blood for ABO Grouping and ABO Subgrouping  
Type Blood of Donors and Receipients  
Test Blood for RHO or DU Factors  
Store Blood According to Grouping and Factor  
Centrifuge and Separate Serum from Clot  
Prepare Blood for Shipment  
Maintain Files of Blood Banking Forms  
Perform Direct and Indirect Coombs Tests  
Record Information on Blood Record Card  
Prepare and Process Specimens  
Heterophile Presumptive and Differential Antibody Test

#### Blood Bank Job Type (Cont'd)

Collect Blood Specimens Directly from Patients  
Dispose of Blood after Time Limit  
Perform Cardiolipin Microflocculation  
Perform C Reactive Protein Tests  
Perform Latex Fixation Test  
Log Incoming or Outgoing Specimens  
Draw Blood for Transfusions  
Maintain Donor Files  
Process Blood for Packed Cells

#### Hematology Supervisor Job Type

Perform Hematocrit Tests  
Perform Blood Count  
Prepare Blood Smears  
Perform Erythrocyte Sedimentation Rate  
Identify Morphological Variations of Blood Cells  
Perform Reticulocyte Count  
Perform Sick Cell Preparations  
Separate Serum from Blood  
Identify Immature Blood Cells  
Perform Eosinophile Counts  
Determine Coagulation Times by Lee White Method  
Perform Spinal Fluid Cell Counts  
Requisition Supplies and Equipment  
Perform Thrombocyte Count  
Determine Coagulation Times by Capillary Method  
Determine Bleeding Time Ivy Method  
Collect Blood Specimens Directly from Patients  
Perform Differential Cell Counts  
Perform Clot Retraction Test  
Determine Bleeding Time Duke Method  
Perform Cerebrospinal Fluid Count  
Perform Erythrocyte Indices

#### Bacteriology Job Type

Prepare Culture Media  
Clean Area and Equipment Aseptically  
Identify and Classify Pathogenic Bacteria  
Perform Antibiotic Sensitivity Test  
Stain Bacteriological Smears  
Examine Specimens Microscopically

#### Bacteriology Job Type (Cont'd)

Identify Protozoans Cestodes Nematodes or Trematodes  
Collect Skin Specimens Directly from Patients  
Perform Concentration and Flotation Techniques  
Collect Pus Specimens Directly from Patients  
Log Incoming or Outgoing Specimens  
Perform Bacteriological or Chemical Exam of Water  
Collect Fecal or Urine Specimens from Patients  
Stain Mycology Specimens  
Stain Parasitological Smears  
Prepare Solutions and Standards  
Maintain Files of Laboratory Records and Reports  
Investigate Possible Sources of Staphylococcus Outbreaks  
Perform Sperm Counts  
Cultivate Mycology Specimens for Primary Isolation  
Perform KOH Preparation for Dermatophytes  
Identify and Classify Fungi  
Collect Sputum Specimens Directly from Patients

#### Histopathology Technician Job Type

Section Tissue in Microscopic Blocks  
Mount Tissue Section in Preparation for Microscopic Study  
Embed Tissue in Paraffin  
Stain Specimens for Microscopic Study  
Prepare Routine Stains  
Prepare Tissue for Dehydration and Infiltration of Paraffin  
Assist with Autopsy  
Prepare Special Stains  
Log Incoming or Outgoing Specimens  
Use Autotechnicon  
Prepare and Process Specimens  
Decalcify Specimens of Teeth and Bone  
Prepare Specimens for Shipment  
Submit Tissue Specimens to AFIP or Histopathology Centers  
Prepare Frozen Section of Tissue  
Use Microtome Knife Sharpener  
Clean Area and Equipment Aseptically  
Collect Biopsy or Autopsy Specimens

#### NCOIC Job Type

Evaluate Work Performance of Subordinates  
Resolve Technical Problems of Subordinates  
Assure the Availability of Equipment and Supplies

NCOIC Job Type (Cont'd)

Assign Specific Work to Individuals  
Evaluate the Accuracy of Routine Reports  
Develop and Improve Work Methods and Procedures  
Plan Reports for the Section  
Plan and Schedule Work Assignments  
Direct Maint Utilzn of Equip Supplies and Work Space  
Determine Equipment Repairs of Replacements Needed  
Evaluate Compliance with Established Work Standards  
Supervise On-the-Job Training Programs  
Evaluate the Adequacy of Routing Reports  
Coordinate Work Activities with other Sections  
Establish Work Priorities  
Show How Locate and Interp Technical Information  
Assist Officer in Charge Estab Organizational Policy  
Evaluate Individuals for Promotions and Upgrading  
Recommend Special Corrective Action for Recurring Problems  
Rotate Duty Assignments of Personnel

Figure 1. 90450 Medical Laboratory Technician

Task Information		% of Group Performing	Ave % Time by Those Performing	Ave % Time by Total Group	Cumulative Sum Ave % Time by all Members
F	18 Collect Blood specimens directly from patients	93.40	1.70	1.58	1.58
J	3 Perform Blood count	89.09	1.56	1.39	2.98
J	17 Perform Hematology procedures for differential cell counts	88.83	1.49	1.33	4.30
J	24 Perform Hematology procedures for Hematocrit tests	89.09	1.46	1.30	5.60
N	2 Examine Urine specimens microscopically	88.07	1.43	1.26	6.85
J	5 Prepare Blood Smears	89.85	1.39	1.25	8.10
F	10 Prepare and process specimens	87.56	1.39	1.22	9.32
N	9 Perform Urinalyses for glucose tests	87.82	1.38	1.21	10.53
N	15 Perform Urinalyses for specific gravity tests	87.06	1.38	1.20	11.74
N	6 Perform Urinalyses for albumin tests	87.06	1.36	1.19	12.92
J	3 Clean area and equipment aseptically	80.96	1.46	1.18	14.11
N	1 Examine Urine specimens macroscopically	87.82	1.32	1.16	15.27
J	6 Separate Serum from blood	87.31	1.30	1.14	16.40
F	11 Prepare reagents	93.40	1.19	1.11	17.52
J	2 Identify morphological variations of blood cells	88.07	1.21	1.06	18.58
M	4 Operate spectro-photometer	77.66	1.34	1.04	19.62
J	21 Perform Hematology Procedures for erythrocyte microfloculation	87.56	1.19	1.04	20.66
K	7 Perform serological procedures for cardiolipin	78.93	1.30	1.03	21.69
G	1 Examine specimens microscopically	86.04	1.18	1.01	22.70
G	2 Identify and classify pathogenic bacteria	78.68	1.27	1.00	23.70
G	10 Prepare culture media	78.68	1.26	0.99	24.69
F	12 Prepare solutions and standards	86.55	1.09	0.94	25.63
M	25 Perform biochemical procedures for liver function tests	78.93	1.18	0.93	26.56
G	11 Stain bacteriological smears	85.28	1.08	0.92	28.42

Figure 2. Data Illustrating Individual Job Descriptions,  
Common Worktime, Consolidated Job Description, and  
Error Values

TASK NR.	% TIME JOB "A"	% TIME JOB "B"	COMMON WORKTIME	CONSOLIDATED DESCRIPTION	ERROR FOR JOB "A"	ERROR FOR JOB "B"
1	20	30	20	25	05	05
2	40	30	30	35	05	05
3	20	20	20	20	00	00
4	20	10	10	15	05	05
5	00	10	00	05	05	05
	<u>100</u>	<u>100</u>	<u>80</u>	<u>100</u>	<u>20</u>	<u>20</u>

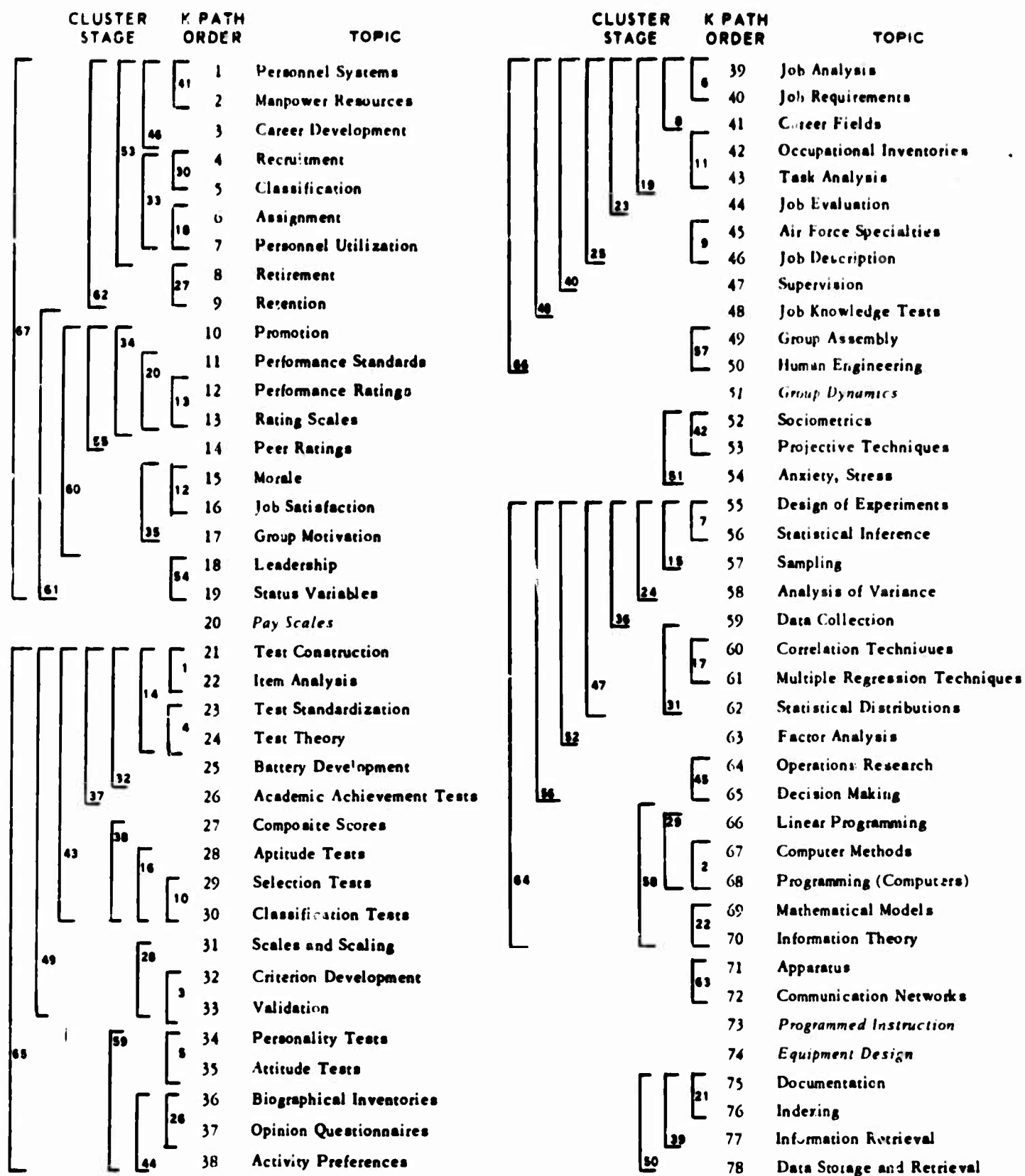


Fig. 3. Clustering of information topics through 67 computer stages. Ungrouped topics in *italics*.



#### FOOTNOTE

The "MAXOF Cluster Model" has been available and utilized for more than five years. However, this is the first time that the model has been given a name. It has on occasion been referred to as "The Personnel Research Laboratory Hierarchical Grouping Model." In other instances, applications of the model have been given a title, such as "The PRL Job-Type Analysis Program" or "The Iterative Criterion Clustering Program." None of these titles is descriptive or easy to remember. Hopefully, future papers will be consistent in applying the name given in this paper, so as to avoid further confusion of the readers.

## THE MULTIVARIATE ANALYSIS OF QUALITATIVE DATA<sup>1,2</sup>

James C. Lingoes  
The University of Michigan<sup>3</sup>

Although a large number of linear techniques have been proposed for the multivariate treatment of quantitative data (Ball, 1965), little has been advanced for the multidimensional analysis of nominal data. Indeed, in some quarters (Torgerson, 1958), nominal or classificatory variables do not merit the status of scales and are, therefore, not deserving of any serious consideration in a book on scaling theory and methods.

When an investigator is confronted with categorical variables in the context of their more respectable brethren, quantitative variables, and he is, nevertheless, determined to analyze them, typically he resorts to constructing "dummy variables" from the various categories before proceeding with a standard linear analysis. Alternatively, the researcher may eliminate nominal variables from the analysis proper (save, perhaps, for that omnipresent dichotomy of sex), resting content to later use the categorical data in a descriptive capacity for talking about his results and interpretations. Each of these strategies has its associated dangers. In brief, for the former choice, an element of artificiality is introduced along with all the problems attendant upon having uneven or extreme marginal distributions - factors which may cause difficulties in interpretation and result in a loss of parsimony. In the latter choice, a very real risk may be incurred in assigning appropriate weights to the qualitative data because of their univariate treatment and, as a consequence, some important cues may be lost for refinement of research design.

This paper will be concerned with presenting the rationale and details of three possible approaches to the multivariate analysis of nominal data. The

1. An invited paper presented at the Conference on Cluster Analysis of Multivariate Data, held in New Orleans, Louisiana on 12/9-11/66.
2. This research in nonmetric methods is supported in part by a grant from the National Science Foundation (GS-929).
3. Prepared while on leave to The University of California, Berkeley.

## 2-Lingoes

three procedures to be discussed in turn are: 1) Multivariate Analysis of Contingencies - II (Guttman, 1941; Lingoes, 1963b; 1964); 2) Multidimensional Scalogram Analysis - I (Guttman<sup>4</sup>; Lingoes, 1966a); and, 3) Multidimensional Scalogram Analysis - II (Guttman, 1967; Lingoes, 1967c), which, for brevity, are labeled: MAC-II, MSA-I, and MSA-II, respectively.

### Multivariate Analysis of Contingencies

The general problem of multidimensional analysis is concerned with the three basic facets of persons (P), variables (J), and categories ( $C_j$ ), the subscript on C always being implied, when not expressly written, to denote that a category belongs to an item, i.e., it has no independent status. Given these sets, our task is one of mapping P into  $C_j$  ( $j \in J$ ) or, symbolically,  $P \rightarrow C_j$ . The characteristic function of the three sets is:

$$1) \quad e_{pjc} = \begin{cases} 1, & \text{if } p \rightarrow c \text{ for } j \\ 0, & \text{otherwise} \end{cases}$$

The binary matrix defined by 1) is called the attribute or trait matrix E, representing the general model for both quantitative and qualitative data. Under the condition of having mutually exclusive and exhaustive categories (which can always be effected by a proper choice or definition):

$$\begin{aligned} 2) \quad \sum_{c \in C_j} e_{pjc} &= 1 \quad (j \in J; p \in P); \\ 3) \quad \sum_{j \in J} \sum_{c \in C_j} e_{pjc} &= n, \text{ the number of variables in set } J \text{ } (p \in P); \\ 4) \quad \sum_{c \in C_j} \sum_{p \in P} e_{pjc} &= N, \text{ the number of persons in set } P \text{ } (j \in J); \end{aligned}$$

-----  
4. Guttman, L. Unpublished lectures on multidimensional analysis given at The University of Michigan, 1965. Much in the above and following discussion is based upon these lectures. Some minor changes in notation and terminology have been made to conform with a prior presentation (Lingoes, 1963b).

- 5)  $\sum_{c \in C_j} 1 = k_j$ , the number of categories in  $C_j$  ( $j \in J$ ;  $p \in P$ );
- 6)  $\sum_{j \in J} k_j = n_c$ , the number of categories over all items; and
- 7)  $\sum_{p \in P} e_{pjc} = n_{jc}$ , or the number of persons in category  $c$  of item  $j$ .

$[\pi_{jc} = n_{jc}/N = \sum_{p \in P} e_{pjc}]$ , or the probability of variable  $j$  falling in category  $c$  = the relative frequency = the expected value of  $e_{pjc}$  for the population  $P$ .]

A universal property of the characteristic function is that any scoring scheme for persons can be formulated in terms of the product of a set of weights and the elements of  $E$ :

$$8) \sum_{j \in J} \sum_{c \in C_j} w_{jc} e_{pjc} = s_p \quad (p \in P).$$

An example of a simple scoring system would be a vector of 1s and 0s for dichotomous items, e.g., the number of correct answers. A more complicated scoring system might make adjustments for guessing, etc. In all cases, however, individuals are placed into score classes such that one person or group of persons is distinguishable from another person or group. The scoring problem can thus be seen as being equivalent to that of determining the partitions of  $P$  under specified constraints. What aspects of  $E$  are we interested in classifying? The answer to this question will specify the kinds of constraints needed for finding a solution to the unknown weights and scores. Some may be interested in the principal components of scales (Guttman, 1950); others in scale homogeneity (Dempsey, 1963); some in optimizing discriminant functions (Bryan, 1961); yet others in reducing dimensionality within the framework of common factor theory (Butler, et al, 1963); and others in maximizing linearity as a basis for both typing objects and determining a smallest space nonmetric solution for variables (Lingoes, 1963b; 1964; 1966b). Again, other interests, as in the MSA-I and MSA-II approaches, will suggest alternative restrictions on the solutions to the partitioning problem. All, however, are

#### 4-Lingoes

based upon E and all, with the exception of MSA-II, use the basic theory and equations worked out by Guttman (1941) a quarter of a century ago.

#### The MAC-II Basic Equations

Matrix notation will be used for outlining the initial steps of the MAC-II algorithm.

STEP 1: Define an  $n_c$ -order diagonal matrix F, whose  $f_{jj}$  elements = the number of Ss falling in the  $j^{\text{th}}$  category, i.e.,  $n_{jc}$ . If there are n variables and N Ss then  $\text{tr}(F) = nN$ .

STEP 2: Form the matrix product  $F^{-\frac{1}{2}}E = G$ , where E and G are  $n_c \times N$  order matrices.

STEP 3: Compute  $G'G = M$ , where M is an N-square Gramian matrix with typical element:

$$m_{pq} = \sum_{j \in J} \sum_{c \in C_j} \frac{e_{pjc} e_{qjc}}{n_{jc}} \quad (p, q \in P; \quad 0 \leq m_{pq} \leq n),$$

or the number of categories that p and q share weighted inversely by the number of persons falling in the shared categories.  $\text{Tr}(M) = n_c$  and the maximum rank of M,  $\rho(M) = n_c - n + 1$  (for  $N \geq n$ ).

STEP 4: Solve the eigenequation:  $(M - \lambda I)S = 0$ , where I is an order N identity matrix and  $\lambda$  and S, respectively, are the roots and vectors satisfying the equation. The largest root of the solution = n and the elements of its associated vector of unit length =  $(1/N)^{\frac{1}{2}}$ , a constant, representing the substantively trivial but formally important solution corresponding to removing "chance expectation" as in chisquare analysis. Indeed, what we are solving for are the orthogonal components of chisquare and the resultant metric is that of chisquare. The mean score for each independent score vector = 0 as a consequence of the con-

5-Lingoes

stant solution.

Although this paper will not be concerned with determining optimal category weights (the mean of the distribution of scores for persons falling in various categories), which serve as the basis for the nonmetric factor analysis of variables by SSA-III (Lingoes, 1966b; Lingoes & Guttman, in preparation) in the later part of the MAC-II program, the significance of the roots should be commented upon. If we order the roots of  $M$  (excluding  $\lambda_{\max}$ ) as follows:  $\lambda_1 > \lambda_2 > \dots > \lambda_j > \dots > \lambda_\rho$ , then the total chisquare can be obtained from:

$$9) \chi^2_t = \frac{N}{2} \sum_{j=1}^{\rho} (\lambda_{j-1})^2, \text{ having } \rho(N-1) \text{ or } (n_c - n)(N-1) \text{ degrees of}$$

freedom. On the other hand, the  $j^{\text{th}}$  partition of chisquare:

$$10) \chi^2_j = \frac{N}{2} (\lambda_{j-1})^2 \text{ and has } [(\rho-1) + (N-1) - 2(j-1)] \text{ degrees of free-}$$

dom. Corresponding to each root, however, there is a correlation ratio:

$$11) \eta_j = (\lambda_{j-1})/(n-1), \text{ which varies between 0 and 1 and measures the}$$

covariance among variables. If all bivariate regressions are linear or can be made linear by finding an optimal set of scores (in the least squares sense of MAC-II), then the correlation ratio will equal the average intercorrelation among the variables. This fact, of course, is exploited in the MAC series for linearizing (Lingoes, 1964) relationships among quantitative variables, e.g., when nonlinear relationships may be present.

By introducing the concept of statistical significance we are afforded a rationale for attending to but a subset of the  $\rho$  vectors for further analysis. Let  $m$  = the number of significant roots in our solution. Each person then can be

## 6-Lingoes

plotted in a  $m$ -dimensional Euclidean space and our problem reduces to that of determining the further partitions of this subspace in terms of salient clusters. Although a number of techniques could be used for clustering, once the proper subspace has been determined, a hierarchical clustering method, max-min cluster analysis (Lingoes, 1963b), based upon a perceptual and statistical model was used and is described below.

### Max-min Clustering

STEP 1: Normalize the unit length vectors of scores to the length of their associated roots, i.e., form  $H^{\frac{1}{2}}S' = X'$ , where  $H$  is the  $m$ -square diagonal matrix of etas and  $S$  and  $X$  are the order  $N \times m$  matrices of unit length score vectors and normalized scores, respectively.

STEP 2: Calculate the  $N$ -square matrix of Euclidean distances among the  $N(N-1)/2$  pairs of persons according to the standard distance formula:

$$12) \quad d_{pq} = \left[ \sum_{i=1}^m (x_{pi} - x_{qi})^2 \right]^{\frac{1}{2}} \quad (p=1,2,\dots,N-1; q=p+1,p+2,\dots,N).$$

STEP 3: Compute  $\bar{d}$  and  $s_d$  from the off-diagonal elements of the distance matrix  $D$ . Set Level,  $l = 0$ ; compute  $\Delta = s_d/8$ ; and set Radius,  $r^{(l)} = s_d/2 - \Delta$ .

STEP 4. Set  $l = l+1$  and  $r^{(l)} = r^{(l-1)} + \Delta$ . If  $l = 14$ , or if  $l = 6$  and less than  $\frac{1}{4}N$  has been clustered, or if the number of clusters is less than 3, terminate clustering. Otherwise proceed to next step.

STEP 5: Sitting on each point in turn in the  $m$ -dimensional space determine the number of points which are within the current radius criterion for each. Select that point accounting for the most points within the radius as a new cluster (breaking any ties in favor of that point having the smallest sum of squared distances between the centroid of the cluster and those points within its orbit).

## 7-Lingo

If any point represents a cluster of persons, then the number of individuals in that cluster are considered whenever a new cluster is to be formed.

Excluding points already classified at a fixed level, determine that point which accounts for the next largest number of individuals, iteratively, until no further clusters can be formed at the given criterion. Compute the centroids of all clusters formed within the radius of inclusion and the reduced matrix of interpoint distances. Go to STEP 4.

Some useful statistics that might be calculated at each level are: 1) the mean and standard deviation of the off-diagonal distances; and 2) for each cluster comprising four or more individuals: a) the mean and variance of the distances from the origin for constituents; b) the interpoint distances between all such pairs of clusters; and c) t-tests (based upon pooled estimates of variance) between pairs of clusters.

The above rather simple clustering procedure results in a tree where Ss are classified in only one cluster at a given level and are never reclassified on the basis that the cluster might thereby be improved. Based upon the statistics computed for each level in conjunction with extra-statistical considerations (e.g., number of clusters desired, number of Ss unclassified, meaningfulness, etc.), the investigator is free to select that level of partitioning most appropriate to his purposes. Indeed, the hierarchical approach has as one of its chief virtues this kind of freedom of selection. Among the statistics calculated at each level, the  $\bar{d}$  and  $s_d$  of the interpoint distances have often proved helpful as guides. Generally, as radius increases so does mean distance among points, but not uniformly nor for all problems at the same level. On the other hand, the standard deviation does not behave as consistently as a function of level. For most purposes choosing that level for which the coefficient of variation is a minimum has proven optimal, i.e.,  $V = 100s_d/\bar{d}$ .



## 8-Lingoes

In summary, starting with a completely general binary matrix  $E$ , representing a subject by category classification, for any set of categories (quantitative and/or qualitative), a solution for a set of real numbers (chisquare metric) is sought to replace the arbitrary category captions such that the covariance matrix among variables is maximized. We impose the restrictions that each scoring system be orthogonal to every other, each successively accounts for the maximum amount of remaining variance, and the number of such scoring systems be a minimum. Since we are not interested in exactly reproducing  $E$ , something less than a full set of vectors are required. To the extent that sampling and measurement errors may have entered into the determination of  $E$ , we have invoked the statistical concept of significance for selecting the appropriate subset of solutions. Finally, as a way of looking at and organizing the configuration of points in  $m$ -dimensional space, we have introduced a clustering procedure whereby each point appears within a set of spherical envelopes of varying radii such that mutually exclusive sets of these spheres define a typology and give us a feel for the distribution of points free of considerations in respect to the origin, rotation, or orientation of the principal axes. Thus, three kinds of partitioning are involved in MAC-II: 1) at the point where the number and kinds of categories have been decided upon (mainly a psychological problem), 2) at the level where the category captions have been replaced by a set of optimal weights (a problem of numerical analysis), and 3) at the level where a subspace is defined and clusters are sought (a problem involving statistical, perceptual, and psychological considerations). The first and third partitions involve components of subjectivity and arbitrariness, while the second is completely objective and results in a unique solution (given the initial set of categories and the function that is being maximized).

Although most applications of MAC have been restricted to quantitative data, some have involved combinations of quantitative and qualitative variables and some

predominantly qualitative data (McPherron, 1963), yielding most interesting results. More studies, however, are required to properly assess the potentials and limitations of this technique for the multidimensional treatment of categorical data.

A redefinition of our goals and the kinds of constraints imposed suggests yet another way of looking at the basic data matrix E.

### Multidimensional Scalogram Analysis - I

The essential task set for MSA-I is that given the  $N$  points embedded in a subspace defined by the  $m$  largest vectors of  $X$  (the normalized score vectors), can we transform the coordinates such that for a fixed item all individuals falling within a given category will be placed in a contiguous region of that space? We are thus seeking a definition of category boundaries yielding regions of indefinite contours (the nature of the boundaries are not specified) where each item represents a partitioning of the space. In order to solve this problem we need to specify how the boundaries are to be determined such that contiguous regions are insured and, further, what is the nature of the loss function to be minimized, i.e., how are we to evaluate noise?

Consider a given partition  $j \in J$  of the  $m$ -dimensional subspace defined by  $X$ , the points falling within a specified category ( $c \in C_j$ ,  $j \in J$ ) will not, in general, fall within a region all of whose members belong to that category. For each point not belonging to  $c$ , however, say,  $b \in C_j$ , there is a closest point that does belong to the category  $c$ ; such a closest point is defined as a trial "outer-point" of category  $c$ . For each person-point in turn we can define a set of outer-points. Further, all points not classifiable as outer-points will be considered as "inner-points". Now, the set of points falling in a fixed category, outer- and inner-points alike, are defined as being contiguous iff each (if any) inner-point is closer to some outer-point of the same category than it is to any outer-point of

## 10-Lingoes

of any alternative category of the same item. More formally, given item  $j \in J$  and  $c \in C_j$  and any three points  $p, q, r \in P$ , if  $e_{pjc}e_{qjc}(1-e_{rjc}) = 1$  and  $d_{pr}^2 \leq d_{qr}^2$ , then  $\alpha_{pj} = 1$  ( $p$  is an outer-point of  $c$  for  $j$ ), otherwise  $\alpha_{pj} = 0$  ( $p$  is an inner-point of  $c$  for  $j$ ), where  $d_{pr}^2$  is the squared Euclidean distance between the two points,  $p$  and  $r$ . The remaining algebra is concerned with an explicit statement of the function to be maximized (Guttman's coefficient of contiguity,  $\lambda$ ), how the coordinates are to be modified in order that the foregoing function is maximized, and how we are to control or modulate the convergence process. The MSA-I program (Lingoes, 1966a) is completely adequate for the analysis of quantitative as well as qualitative data, dichotomous as well as n-chotomous variables, monotone and/or polytone items, and involves no assumptions whatsoever about scaling properties or distributions.

### The MSA-I Basic Equations

STEP 1: Determine what outer-point of its own category is  $p$  as an inner-point closest to, i.e., if  $e_{pjc}(1-\alpha_{pj})\alpha_{qj}\alpha_{rj} = 1$  and  $d_{pq}^2 \leq d_{pr}^2$ , then  $\beta_{pqj} = 1$  and otherwise  $= 0$ . If  $\alpha_{pj} = 1$ , then  $\beta_{pqj} = 0$  for all  $q$ .

STEP 2: Determine what outer-point of another category is  $p$  as an inner-point closest to, i.e., if  $e_{pjc}(1-\alpha_{pj})(1-e_{qjc})\alpha_{qj}(1-e_{rjc})\alpha_{rj} = 1$  and  $d_{pq}^2 \leq d_{pr}^2$ , then  $\gamma_{pqj} = 1$ , otherwise 0.

STEP 3: Where  $\beta_{pqj}$  is a  $n$  element column vector and  $\gamma_{pqj}$  is a  $n$  element row vector compute:  $\epsilon_{pqr} = \sum_{j=1}^n \beta_{pqj}\gamma_{prj}$ .

STEP 4: Compute the sign matrix:  $S_{pqr} = \text{sgn}(d_{pr}^2 - d_{pq}^2) = -S_{prq}$ .

STEP 5: Calculate:  $\epsilon_{pqr}^* = S_{pqr} \cdot \epsilon_{pqr}$ , i.e., modify  $\epsilon_{pqr}$  according to

whether the sign of the difference between the squared distances is  $+$  or  $-$ .

# 11-Lingoes

STEP 6: Compute:  $n_{pq} = \sum_{r=1}^N f_r (\xi_{prq} - \xi_{pqr})$  and  $n^*_{pq} = \sum_{r=1}^N f_r (\xi^*_{prq} - \xi^*_{pqr})$ ,

where  $f_r$  = the number of persons in the  $r^{th}$  type, i.e., individuals having identical profiles over the  $n$  variables, and  $N$  = the number of types rather than persons. (N.B. If each of the points is to be weighted by this  $N$  element frequency vector, then the initial configuration based on the  $N$  types should be adjusted so that the weighted mean of  $X$  is zero for each vector. Letting  $U$  represent the unweighted normalized score vectors, then:  $X_{pa} = U_{pa} - \sum_{r=1}^N f_r U_{ra} / N$  ( $p=1,2,\dots,N$ ;  $a=1,2,\dots,m$ ).  $X$  will now be referred to as the weighted normalized score matrix.)

STEP 7: Calculate the  $N$ -square matrix  $M$  with typical off-diagonal element:  $m_{pq} = -(n_{pq} + n_{qp})$ ,  $q \neq p$  and typical diagonal element:  $m_{pp} = \frac{-1}{f_p} \sum_{q=1}^N f_q m_{pq}$ ,  $q \neq p$ . The row and column sums of  $M = 0$ .

STEP 8: Similarly compute  $M^*$  by substituting  $n^*_{pq}$  for  $n_{pq}$  and  $n^*_{qp}$  for  $n_{qp}$ .  $M^*$  will also be an  $N$ -square matrix whose rows and columns sum to zero.

STEP 9: Calculate:  $W_{pa} = f_p X_{pa}$ .

STEP 10: Determine:  $\lambda = \frac{\sum_{a=1}^m W'_a M W_a}{\sum_{a=1}^m W'_a M^* W_a}$ , the coefficient of contiguity, which

varies between  $-1$  (representing perfect discontiguity) and  $+1$  (representing perfect contiguity).

STEP 11: If  $t$  (the number of iterations, initially set = 1) = some preset number or if  $\lambda$  = some predetermined cut-off point, then increase  $m$  to  $m+1$  and reset  $t=1$ , provided that more dimensions are required to get a good fit, otherwise terminate. When going to a higher dimensionality one always starts with the initial configuration in order that the metric be comparable from one set of dimensions to that which is appended. If neither of the first two conditions obtain, go to the next section for computing a new trial set of coordinates

## 12-Lingoes

STEP 12: Calculate the Nom matrix:  $Y_{pa} = X_{pa}F(M - \lambda M^*)$ , as a basis for modifying X.

STEP 13: Compute:  $\beta = \sum_{a=1}^m \sum_{p=1}^N f_p Y_{pa}^2$ .

STEP 14: If  $t=1$  calculate the scalar  $c = \frac{1}{t}(1 - \lambda)$ , otherwise:  
 $c = \beta^{(t)} / \beta^{(t-1)}$  if  $c \leq 1$ , otherwise set  $c = 1$ .

STEP 15: If  $t=1$  compute the scalar:  $\sigma = \frac{1}{2}(1 - \lambda)$ , otherwise:  
 $\sigma = c^{(t-1)} \sigma^{(t-1)}$ .

STEP 16: Also compute the scalar:  $\alpha = \sum_{a=1}^m \sum_{p=1}^N f_p X_{pa}^2$ .

STEP 17: Compute the multiplicative scalar which keeps the results of adjacent iterations highly correlated:

$k = [(\alpha c \sigma) / (\beta(1 - (c \sigma)))]^{\frac{1}{2}}$ , a value which is a monotonic decreasing function of  $t$  and never reaches zero unless and until  $\lambda = 1$ .

STEP 18: Compute the new set of coordinates:  $Z = X + kY$ .

STEP 19: Set new coordinates equal to the initial squared Euclidean norm:  
 $X^{(t+1)} = Z [(\sum_a \sum_p f_p X_{pa}^{(1)2}) / (\sum_a \sum_p f_p Z_{pa}^2)]^{\frac{1}{2}}$ , ( $p=1,2,\dots,N$ ;  $a=1,2,\dots,m$ ).

STEP 20: Set  $t=t+1$ , compute the matrix of squared Euclidean distances, and return to STEP 1.

The above "average steepest ascent" algorithm in general converges in a few iterations, but is a time-consuming process in that each cell of  $M$  and  $M^*$  involves  $N(N-1)/2$  calculations and each of these in turn are based on a large number of computations involving the  $n$  items and the  $n_c$  categories. MSA-I's complete generality for quantitative and/or qualitative data and for linear and/or nonlinear relationships makes it an ideal procedure for studying both numeric and conceptual problems (e.g., facet models: Guttman, 1959). Although the solutions resulting from MSA-I are embedded in an Euclidean space, the dimensions of this space are

### 13-Lingoes

not, in general, meaningful. One must look at the configuration of points in this space and study each partition separately in terms of the properties of regions in order to make full use of this method. Since neither rectilinear nor parallel boundaries are insisted upon in the definition of contiguity, one loses potential information in respect to order for both items and categories. Indeed, with its weak definition of contiguity the method often results in what might be considered a quasi-topological representation - very revealing and certainly fascinating from a number of points of view.

In summary, starting with the basic data matrix E and defining a trial space based upon the normalized weighted score matrix X, the task set for MSA-I is that of moving the points around in this space such that a certain definition of contiguity is satisfied in a minimum number of dimensions. Each type is a point in Euclidean space, each item is a partitioning of this space, and each region within a partition represents a category. A subset of the person points, i.e., those characterized as "outer-points", define the contiguous regions which may assume any form whatsoever. The cutting points of Guttman's earlier technique of scalogram analysis (1944) for  $m=1$  lie between the outer-points of MSA-I. When  $m=2$  there are cutting curves and for  $m \geq 2$  there are cutting surfaces separating the boundary definers. Contrasted with the earlier method of scale analysis, not only is the number of errors counted (as reflected in the coefficient of reproducibility), but the size of the errors is also taken into consideration by the coefficient of contiguity. For example, take a variable like religion whose three categories were: a = catholic, b = protestant, and c = jewish and a particular individual p who fell in category a. Now if  $d_{pr}^2$  was the smallest squared distance of p from all other points and r happened to define the category boundary of protestants (r is an outer-point for category b), then a decrement to the coefficient of

contiguity would be incurred.

As an illustration of MSA-I the following example of a purely conceptual analysis is given.

#### An MSA-I of Social Structure

Guttman's adaptation (1966 ) of a table appearing in Bell and Sirjamaki's (1961; p. 325) sociology text provides a set of five characterizations by which groups of persons can be differentiated. The five facets and their elements or categories are as follows: 1) Intensity of Interaction (a = slight, b = low, c = moderate, and d = high); 2) Frequency of Interaction (a = slight, b = non-recurring, c = infrequent, and d = frequent); 3) Feeling of Belonging (a = none, b = slight, c = variable, and d = high); 4) Physical Proximity (a = distant and b = close); and 5) Formality of Relationship (a = no relationship, b = formal, and c = informal). The objects to be classified, seven in number, are various kinds of groups, i.e.: 1) Crowd (aaabb); 2) Audience (bbbbb); 3) Public (aabaa); 4) Mob (dbdbc); 5) Primary Group (dddbc); 6) Secondary Group (cccab); and 7) Modern Community (bccbb). With no a priori conceptions as to order in respect to types of groups, items, or categories within an item the following perfect solution in two dimensions required but one iteration (see Fig. 1).

- - - - -

Figure 1 about here

- - - - -

The rather interesting Y-configuration that emerges places Primary Group and Mob close together and at the foot of the Y and the remaining groups in the order: Secondary Group, Modern Community, Audience, Crowd, and Public - forming the arc. It will also be noted that for each of the five characteristics there appears a

## 15-Lingoes

circular arrangement for the categories for those types appearing on the V part of the Y, e.g., in respect to Intensity of Interaction the ordering goes from moderate to low to slight following the above ordering for the five groups. Since parallel straight line boundaries can be constructed for this configuration, some purchase on item and category orders can be obtained. The items and the category orders within items can be arranged thusly: Physical Proximity (ab), Feeling of Belonging (cbad), Formality of Relationship (cab), then either order of the following two: Intensity of Interaction (cbda), and Frequency of Interaction (cbda), with the first and last items yielding cutting curves which are orthogonal to each other and the intermediate items having slopes for their boundaries that are at a slant. Because there are so few points and a relatively small number of categories, alternative parallel straight line solutions are possible. For example, Guttman's hand solution of this problem yielded the following order for items: 45312, where within each facet the order of the categories was maintained (1966 ). Furthermore, in his analysis he placed Primary Group as being closest to Secondary Group, whereas MSA-I places these two furthest apart. Without belaboring the point, since this is but an example, the differences between the hand solution and the MSA-I solution may have arisen from the ambiguity of the categories within some of the items, e.g., should the order for Frequency of Interaction be: bacd, abcd, bcad, or some other order? We know that the first three are opposed to the last, but there might be some question as to the best order for the first three. Similarly for the categories of the third item. As can be seen from a discussion of these differences, MSA-I may prove fruitful for testing not only certain substantive issues but may also be revealing in respect to preconceived coding assumptions, e.g., that the categories follow a linear order.

We will now pass on to the third and last technique of this paper, MSA-II.



Multidimensional Scalogram Analysis - II

Starting once again with the binary attribute matrix  $E$ , let us formulate a specification that will reproduce  $E$  in the smallest possible space. One way of looking at the basic data matrix would be that  $E$  is an incomplete proximity matrix having but two coefficients, i.e., 1 and 0, where the rows of this matrix represent categories and the columns persons. Thus, whenever a 1 appears we can assume that some category is in a sense near some person and that the relationship is symmetric. In essence, the  $n_c \times N$  rectangular matrix  $E$  can be thought of as a partial adjacency matrix for a graph whose dimensionality we seek. Fortunately, Guttman (1964) has established the necessary theorems for defining the dimensionality of graphs in terms of smallest space theory. We can now write our specification as: given  $E$ , satisfy the inequality that whenever  $e_{pjc} = e_{qjc} - 1 = 1$  then  $d_{pjc} \leq d_{qjc}$  for all  $p \in P$ ,  $c \in C_j$ , and  $j \in J$  such that the loss function, normalized  $\phi$  (v.i.), is minimized for a specified  $m$  dimensions, where  $d$  is the Euclidean distance. We are defining binary relations in terms of a distance function such that all points belonging to one set (categories) which are in relation to points in the other set (persons) will have smaller distances in the joint space of persons and categories than all points (one from each set) which are not in relation, i.e.,  $e_{pjc} = 0$ . Nothing in this statement is implied about the relationships among categories, items, or persons (this information does not exist, although it could be defined in terms of the relationships between rows and columns of  $E$ ) - all that does exist in  $E$  is the category-person relation. By confining our attention to the inter-set relations, however, we should be able to infer something about the intra-set relationships that are implicit in  $E$  from the nature of our solution.

The following is an outline of the MSA-II algorithm (Lingoes, 1967c).

The MSA-II Basic Equations

STEP 1: Define an  $n_c + N = k$ -square symmetric matrix  $V$  whose first  $n_c$  rows (columns) represent the category captions and whose  $n_c + 1$  to  $k$  rows (columns) represent the person captions of  $E$ . The basic data matrix appears as an off-diagonal submatrix of  $V$  occupying rows 1 to  $n_c$  and columns  $n_c + 1$  to  $k$ . For the elements of the submatrix  $E$ , whenever  $e_{pjc} = 1$  substitute  $1 - (Nn + 1)/(k(k - 1))$  and whenever  $e_{pjc} = 0$  substitute  $1 - (N(n + n_c) + 1)/(k(k - 1))$ . All other off-diagonal elements of  $V$  are set equal to  $\frac{1}{2} - (n_c N + 1)/(k(k - 1))$ . The  $k$  diagonal elements of  $V$  are calculated from the following formula:  $v_{ii} = k - \sum_{j=1}^k v_{ij}$  ( $i=1, 2, \dots, k; i \neq j$ ). The row (column) sums of  $V = k$ , the order of the matrix and  $\text{tr}(V) = 1 + k + k(k - 1)/2$ .  $V$  is a Gramian matrix whose largest root  $= k$  and the elements of its associated unit length vector  $= (1/k)^{\frac{1}{2}}$ , a constant.

STEP 2: Solve for:  $U(V - \lambda I) = 0$ , which yields the initial configuration (see: Lingoes, 1967a), where  $I$  is the order  $k$  identity matrix and  $U$  and  $\lambda$ , respectively, are the vectors of unit length and the roots. Normalize the unit length vectors to the size of their associated roots, i.e.,  $X = U\lambda^{\frac{1}{2}}$ . Ignoring the constant vector, order vectors by their length from large to small. The mean of each vector will be zero.

STEP 3: Calculate the  $n_c N$  Euclidean distances between every category point, on the one hand, and every person point, on the other, i.e.,  $d_{ij}$  ( $i=1, 2, \dots, n_c; j=n_c + 1, n_c + 2, \dots, k$ ) based on the  $m$  dimensions of  $X$ , where  $m$  = some predetermined number based upon either a parameter or the number of vectors whose roots are  $\gg k/2$  (excepting the largest root).

STEP 4: Permute the  $Nn$  smallest distances so that they occupy the same positions that the 1s have in  $E$  and permute the remaining  $N(n_c - n)$  distances to the positions in  $E$  occupied by 0s. These cell-wise permuted distances are the rank

18-Lingoes

images (Guttman, 1967 ; Lingoes, 1967a) of the distances, the  $d^*$ 's, which within tied blocks (e.g., the block of 1s) have been ordered from low to high.

STEP 5: Calculate the normalized phi coefficient of monotonicity:

$$\phi = \frac{\sum_i \sum_j (d_{ij} - d^*_{ij})^2}{2 \sum_i \sum_j d_{ij}^2}, \quad (i=1,2,\dots,n_c; j=n_c+1,n_c+2,\dots,k),$$

and the coefficient of alienation:  $\alpha = (1 - (1-\phi)^2)^{\frac{1}{2}}$ , which permits us to gauge how good our fit is in respect to reducing the error of estimate.

STEP 6: Calculate the coefficient of reproducibility:

$$R = 1 - \frac{\text{number of errors}}{2nN}, \text{ where errors are defined as the number of}$$

distances which are smaller than the largest distance for 1s of E whose positions correspond to 0s plus the number of distances which are larger than the smallest distance for 0s of E whose positions correspond to 1s. This measure disregards the magnitudes of the errors implied in the distances, being solely concerned with the number of such incorrect predictions. When  $\phi = 0$  it must be true that all distances between categories and persons for which  $e_{pjc} = 1$  are smaller than distances corresponding to cell entries of E which are zero. We are thus defining the radius of a circle (more generally that of a sphere) such that all points falling within that enclosure are in relation to the point lying at the center. All points lying within the sphere for which  $e_{pjc} = 0$  plus all points lying outside the sphere for which  $e_{pjc} = 1$  are considered errors. R serves no functional purpose in the MSA-II program, but is an interesting descriptive measure telling us how well we could reproduce E.

STEP 7: If we have satisfied a given number of iterations, or if  $\phi$  is sufficiently small, or if  $\phi$  has not changed significantly over a number of itera-

## 19-Lingoes

tions, we can terminate the solution and then go either up or down in dimensionality according to the same options as in SSA-I (Lingoes, 1965 ; 1967a). If, however, none of these conditions prevail, proceed to the next set of steps for modifying the coordinates for another iteration.

STEP 8: For each of the  $d^*$ 's corresponding to the 1s substitute a mean  $d^*$  and for each of the  $d^*$ 's corresponding to the 0s of E substitute their mean  $d^*$ . It can be seen that we are tying all distances that should be tied such that when a solution has been achieved ties will be broken in an optimal fashion.

STEP 9: Define a  $k$ -square symmetric matrix C, the correction matrix, which is coordinate in respect to the partitions of V in STEP 1. Proceeding from top to bottom, and within each, from left to right, let us number these partitions thusly: I, II, III, and IV. The elements of these four partitions of C are:

Partition I of order  $n_c$ :  $c_{ij} = 0$  ( $i \neq j$ ) and  $c_{ii} = n_c + \sum_{j=n_c+1}^k \bar{d}_{ij}^* / d_{ij}$ ;

Partition II of order  $n_c \times N$ :  $c_{ij} = 1 - \bar{d}_{ij}^* / d_{ij}$ ;

Partition III of order  $N \times n_c$ : II' or  $c_{ji}$  of II; and,

Partition IV of order  $N$ :  $c_{ij} = 0$  ( $i \neq j$ ) and  $c_{ii} = N + \sum_{j=1}^{n_c} \bar{d}_{ji}^* / d_{ji}$ .

Each row (column) of C sums to the constant  $k$  and  $c_{ij} = c_{ji}$ . When  $\phi = 0$ , C becomes a scalar matrix.

STEP 10: Compute a new trial set of coordinates by the following pivotal formula:

$$x_{ia}^{(t+1)} = \frac{1}{k} \sum_{j=1}^k x_{ja}^{(t)} c_{ij}, \text{ where } t = \text{iteration number.}$$

STEP 11: Calculate the  $n_c N$  distances based upon the transformed coordinates and go back to STEP 4 for another iteration. As an alternative to doing just one

## 20-Lingoes

least squares adjustment for every rank image permutation, one could do ten least squares corrections for every permutation, as is done in most of the smallest space programs, by returning to STEP 8, after which one would return to STEP 4.

In summary, based on the binary attribute matrix  $E$  and using a trial set of coordinates which are a function of the ranks of the values in  $E$ , we specify an orthogonal solution in a minimum number of dimensions such that for all pairs of categories and persons it will always be true that whenever  $e_{p,jc} = e_{q,jc} - 1 = 1$  then  $d_{p,jc} \leq d_{q,jc}$ . Short of perfect monotonicity, however, for the given dimensionality we will minimize the function,  $\phi$ . When the process of least-squares-rank-image-permutations converges, we attain a representation in a joint Euclidean space of the two sets of points (categories and persons), such that having defined the largest distance of all points which form a binary relation (the 1s of  $E$ ) we are able to draw spherical boundaries, using each point in turn as a center. The radius corresponding to the largest distance will enclose all points in relation to the center point, thus permitting us to reproduce the original response matrix.

Rather than employing the spherical boundaries outlined above, however, one could partition the joint space by finding those hyperplanes which bisect pairs of points within, but not between, items. These hyperplanes would then cut out regions of the space having linear rather than circular boundaries. From either conception one could determine to what extent  $E$  is reproducible. The curved boundary formulation, however, is more easily implemented and produces a less cluttered picture.

The dimensions of the MSA-II solution are primarily meaningful (or some rotation thereof) in terms of the configuration of person points, although under some circumstances where such a configuration allows parallel straight line boundaries, one may gain some insight into item and category structure (v.i.).

21-Lingoes

An MSA-II of Social Structure

For comparison purposes we will present an MSA-II analysis of the same data analyzed by MSA-I involving seven types of social groups defined by five kinds of characteristics. Once again a two space is perfectly adequate to portray all the interrelationships of E (Figure 2 below).

- - - - -

Figure 2 about here

- - - - -

It will be noted that Figure 2 only contains the person points since this aspect interests us mostly. A comparison of the MSA-I and II configurations of the seven groups reveals a remarkable similarity between the two, although the rationale of these two methods differ greatly. A slightly tilted Y is apparent, Primary Group is closest to Mob, and a circular order among the categories is evident for the five items among the groups arrayed on the V portion of the Y. Given the profiles of these seven groups a set of linear parallel boundaries can be constructed such that for each item all individuals falling in a particular category will be contiguous. The partitions of this space (despite the configurational similarities to Figure 1) are different from the results of MSA-I. Thus, the item and the within item orderings are: Physical Proximity (ab), Formality of Relationship (abc), Frequency of Interaction (cbda), Intensity of Interaction (cbad), and Feeling of Belonging (cbad). As was mentioned before, there would appear to be some ambiguity in respect to category ordering within items, permitting alternative solutions. Furthermore, based upon the MAC-II category weights there is a strong suspicion that curvilinear relationships exist among these variables giving rise to the differences noted between Guttman's hand solution

and the MSA-I and II solutions alike.

Although this one example is insufficient for making any inferences about what will happen in general when MSA-I is compared with MSA-II, certain observations having both a practical and theoretical import are relevant.

#### Some Comparisons Between MSA-I and MSA-II

First, in respect to the size of a problem that can be analyzed by these two scalogram programs, MSA-I has a greater capacity (i.e., up to 50 variables, with as many as 20 categories for each, and up to 60 types) than MSA-II, which is restricted to  $n_c + N \leq 80$ . Second, MSA-II, being based upon a much simpler algorithm, is considerably faster than MSA-I for problems of the same magnitude. Third, the simpler contours for the boundaries of MSA-II are more easily depicted and the resulting representation is easier to grasp vis-a-vis the basic data matrix E. Fourth, MSA-II would seem to have more applications than MSA-I, since (with a minor modification) the former is not restricted to mutually exclusive categories. Fifth, in respect to the criterion of reproducibility, MSA-I reserves a subset of the person points for defining regions and these points are not considered in computing the coefficients of either contiguity or reproducibility. In contrast, MSA-II (at the expense of including a set of points for categories) does include all person points in determining reproducibility and, as such, is more analogous to Guttman's original conception of unidimensional scalogram analysis (1944) and Lingoes' generalization thereof for multiple unidimensional scalogram analysis (MSA) for binary data (1960; 1963a).

Both procedures, starting with the same general data matrix E, are ideally suited for the multidimensional analysis of qualitative data and for quantitative data where the distributional and linear assumptions of standard multivariate

techniques cannot be met or are questionable. Based upon quite different definitions and specifications (MSA-I involving a definition of contiguity in terms of outer-points, for example, and MSA-II being based upon the logic of smallest space analysis (see: Lingoes, 1967b for a review) and a definition of distance and dimensionality for graphs (Guttman, 1964)), the two procedures would appear to yield essentially the same results. Further analyses, however, are necessary before reaching a final conclusion on this point. There may well be certain kinds of data which are more economically represented by one procedure than the other.

#### Summary

Three methods for analyzing qualitative data were introduced: 1) Multivariate Analysis of Contingencies - II (based on the early work of Guttman, 1941), 2) Multidimensional Scalogram Analysis - I (involving a unique definition of contiguity which presupposes a minimum of assumptions), and 3) Multidimensional Scalogram Analysis - II (involving a graph theoretic and smallest space logic). An outline of the basic equations and assumptions of each were presented. One example of a conceptual data matrix was analyzed by both MSA-I and MSA-II and the results were discussed vis-a-vis a hand analysis of the same data based upon a linear ordering of the categories involved. Finally, some comparisons between the two scalogram procedures were made.

In conclusion, the three methods discussed in this paper for the multidimensional analysis of both qualitative and quantitative data and of both linear and nonlinear relationships are based upon a minimum number of assumptions (more consonant with our usual ignorance regarding the metric and distributional properties of social science data). One would anticipate, therefore, an increasing use of these and similar techniques (e.g., the various programs developed by Shepard, 1962; Shepard & Carroll, 1966; Shepard & Kruskal, 1964; Kruskal, 1964;



and McGee, 1967). Shepard's 1962 breakthrough paper provided much of the impetus for these current developments in nonmetric methodology (Lingoes, 1967b).

#### References

- Ball, G. H. Data analysis in the social sciences: What about the details? Proc. - Fall Joint Computer Conference, 1965, 533-559.
- Bell, E. H. & Sirjamaki, J. Social Foundations of Human Behavior. Harper, N.Y., 1961.
- Bryan, J. G. Calibration of qualitative or quantitative variables for use in multiple-group discriminant analysis. Travelers Insurance Cos., Hartford, Conn., Scientific Rep., 1961, 2, 1-26.
- Butler, J. M., Rice, L. N., & Wagstaff, A. K. Quantitative Naturalistic Research. Prentice-Hall, N.J., 1963.
- Dempsey, P. The dimensionality of the MMPI clinical scales among normal subjects. J. Consult. Psychol., 1963, 27, 492-497.
- Guttman, L. The quantification of a class of attributes. In P. Horst, et al, The Prediction of Personal Adjustment. Social Science Research Council, N.Y., 1961, 321-347.
- \_\_\_\_\_ A basis for scaling qualitative data. Amer. Sociol. Rev., 1944, 9, 139-150.
- \_\_\_\_\_ The principal components of scale analysis. In S. A. Stouffer, et al, Measurement and Prediction. Princeton University Press, Princeton, 1950, chap. 6.
- \_\_\_\_\_ Introduction to facet design and analysis. In Proc. of the 15th International Congress of Psychology. Brussels, 1959, 130-132.
- \_\_\_\_\_ A definition of dimensionality and distance for graphs. (Unpubl. Mimeo., 1964).
- \_\_\_\_\_ The nonmetric breakthrough for the behavioral sciences. Automatic Data Processing Conference of the Information Processing Association of Israel. Jerusalem, 1966, 1-16.
- \_\_\_\_\_ A general nonmetric technique for finding the smallest Euclidean space for a configuration of points. Psychometrika, 1967, (in press).
- Kruskal, J. B. Multidimensional scaling: a numerical method. Psychometrika, 1964, 29, 1-27.
- \_\_\_\_\_ Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 1964, 29, 115-129.

25-Lingoes

Lingoes, J. C. Multiple Scalogram Analysis: A Generalization of Guttman's Scale Analysis. Unpublished Ph.D. thesis, Michigan State University, 1960.

Multiple Scalogram Analysis: A set-theoretic model for analyzing dichotomous items. Educ. & Psychol. Measmt., 1963, 23, 501-524. (a)

Multivariate analysis of contingencies: An IBM 7090 program for analyzing metric/nonmetric or linear/nonlinear data. Comp. Rep., 1963, 2, 1-24 (University of Michigan). (b)

Simultaneous linear regressions: An IBM 7090 program for analyzing metric/nonmetric or linear/nonlinear data. Behav. Sci., 1964, 9, 87-88.

An IBM 7090 program for Guttman-Lingoes smallest space analysis - I. Behav. Sci., 1965, 10, 183-184.

An IBM 7090 program for Guttman-Lingoes multidimensional scalogram analysis - I. Behav. Sci., 1966, 11, 76-78. (a)

An IBM 7090 program for Guttman-Lingoes smallest space analysis - III. Behav. Sci., 1966, 11, 75-76. (b)

New computer developments in pattern analysis and nonmetric techniques. In Proc. of the IBM Symposium: Computers in Psychological Research. Blaricum. Gauthier-Villars, Paris, 1967. (a)

Recent computational advances in nonmetric methodology for the behavioral sciences. In Proc. of the International Symposium: Mathematical and Computational Methods in Social Sciences. Rome. International Computation Centre, 1967. (b)

An IBM 7090 program for Guttman-Lingoes multidimensional scalogram analysis - II. Behav. Sci., 1967, 12, (in press) (c)

Lingoes, J. C. & Guttman, L. Nonmetric factor analysis by rank reduction. (being submitted for publication).

McGee, V. E. Elastic multidimensional scaling: A hybrid technique. A paper presented at The International Symposium: Mathematical and Computational Methods in Social Sciences. Rome, International Computation Centre, 1967.

McPherron, A. Programming the IBM 7090 for optimizing taxonomy in archeology. A paper presented at the 1963 Convention of the American Association of Archeology, 1963.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. - I. Psychometrika, 1962, 27, 125-140; - II. Psychometrika, 1962, 27, 219-245.

Shepard, R. N. & Carroll, J. D. Parametric representation of nonlinear data structures. A paper presented at the International Symposium on Multivariate Analysis. Academic Press, 1966.

26-Lingoes

Shepard, R. N. & Kruskal, J. B. Nonmetric methods for scaling and for factor analysis. Amer. Psychol., 1964, 12, 557-558.

Torgerson, W. S. Theory and Methods of Scaling. Wiley & Sons, N.Y., 1958.

Figure 1

MSA-I Configuration of Social Groups

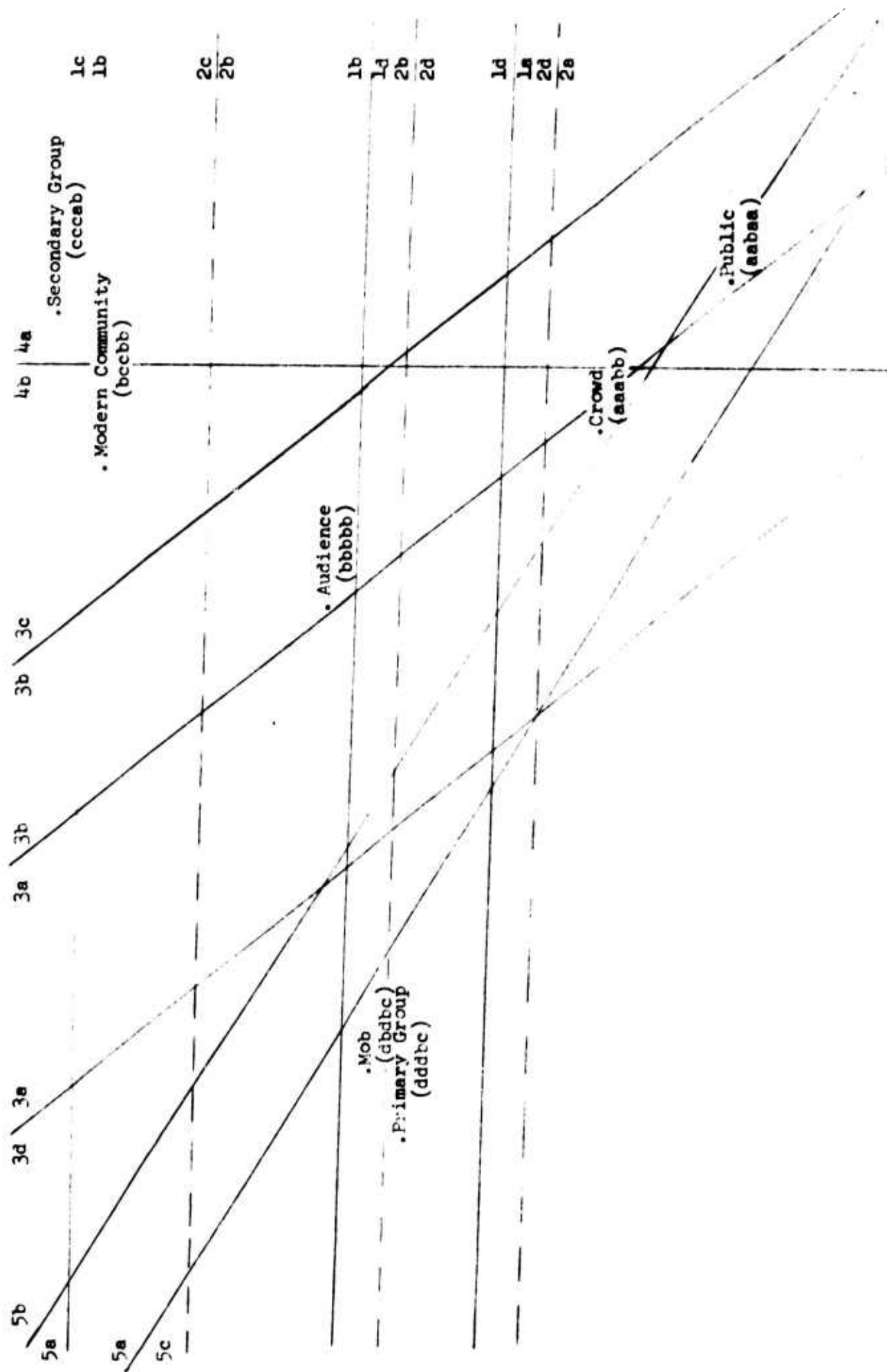
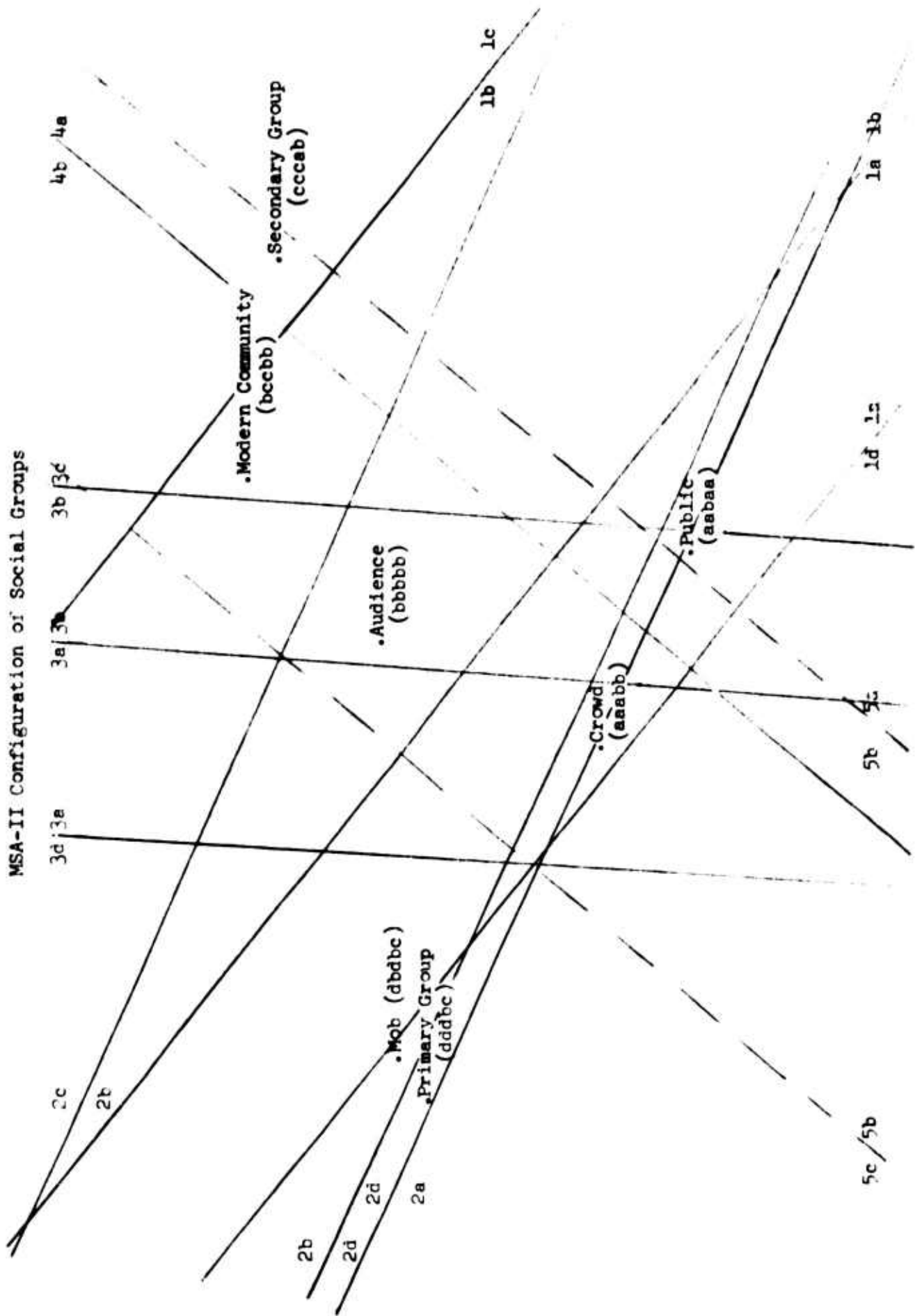


Figure 2

MSA-II Configuration of Social Groups



## "CLASSIFICATION SO AS TO RELATE TO OUTSIDE VARIABLES"

Edward W. Forgy  
UCLA

### Introduction:

Let me first explain a bit of history behind the title. About a year ago in Washington, D.C., there was a conference on classification in psychiatry.<sup>(1)</sup> One of the desirable qualities for any classification system that was unanimously agreed upon by the conferees was that the classes should be relevant to a number of other qualities about persons beyond the information that went into making the classification or diagnosis itself.

At this same conference, a number of empirical papers (2,3,4,5,6) were presented in which typologies were developed from various kinds of data by several computer techniques. In most of these studies, after a classification system was developed, it was then evaluated to some degree by relating it to outside variables not used in building the system.

I couldn't help being struck then by the real absence of any relation between what classifications were hoped to accomplish versus the methods used to develop them. In no case did any outside variables enter in any way into the computing that developed the classification systems. This is analogous to a situation in which relevant linear functions of variables (rather than classifications) were desired, if factor analysis of predictor variables were relied upon exclusively as the method of obtaining the linear functions. Given a choice, most of us would naturally put available criterion variables into the analysis as dependent variables. Then, if such relations are in the data the method will find them, and the linear functions that emerge will be systematically related to the dependent variables of interest. Analogously, I couldn't help thinking that a systematic effort to find relevant classification systems would probably be much more successful than what have been essentially random efforts with respect to relevance.

There are, I think, several kinds of reasons to account for the absence of clustering methods directed at relevance to outside measures. It appears that many investigators, at least psychologists who use cluster analysis, tend to believe that the computer has revealed some sort of natural, pre-existing typological structure in their data. Believing -- or perhaps really only hoping -- this, they consider relevance to particular outside variables a somewhat secondary issue.

The comment by Dr. Sokal on the tendency of various clustering methods to bias the results is most pertinent here. As he pointed out, each method tends to impose a certain kind of structure upon the data, whether that structure is

---

<sup>1</sup>Support received from the National Institute of General Medical Science (GM 11055) and the Health Sciences Computing Facility of UCLA (NIH Grant FR-3) is gratefully acknowledged. The work was done at the Alcoholism Research Clinic, supported by the California Department of Public Health, Division of Alcoholism.

really present or not. I think this applies with even more force to psychological data. In biology, the question is which arrangement of classes fits the data best. In psychology, I think the basic, usually un-asked question is whether any structure of types or classes is called for to describe the data. To help answer this, I'd strongly recommend a simple experiment to anyone who cluster-analyzes his data in order to understand it.

1. Generate some samples of artificial "cases" from a unimodal, non-clustered, non-biased population -- for instance, a joint normal population with the same co-variances as the actual data.
2. Put these through the same cluster analysis process as your real data.

The subtypes that may well be "discovered" in a sample from this classic, classless population would provide a helpful baseline against which to compare the results from the real data.

It is also true that relevance to outside variables is only one of the desirable qualities of a typology, and in addition the most desired kind of relevance is to some broad, usually undefined set of factors, rather than to a particular outside variable. However, it remains that, while some kind of relevance to outside variables is strongly desired for classifications, the main thing depended upon to get it is extraordinary good luck. This being so, I felt that it might be worthwhile to attempt to develop something to improve the odds in favor of an investigator who wants some sort of relevance.

#### The Concept of Maximally Relevant Classes:

Here another acknowledgment is due, in this case a very delayed appreciation of another person's idea. About five years ago a UCLA colleague, Dr. James McQueen, (7), distinguished several possible goals of classification. One of the goals spelled out was so that class membership could be used to estimate or predict some outside variable with minimum error. If the outside variable is a quantitative one, on an interval scale, then the familiar success criterion of minimizing the squared errors can be used. Then the most relevant classification would be that which explains the largest amount of variance of the dependent variable, just as in linear regression.

This is about the simplest kind of relevance possible -- that class membership itself and alone will enable one to make good predictions about one other characteristic. Because it is so simple, it will be the problem dealt with in this paper.

This is not to deny that broader or more subtle kinds of relevance exist. Two kinds of extensions are obvious:

- 1) Asking for the same kind of relevance but to larger set of variables.

- 2) Asking for a different kind of relevance, i.e. that class membership will sharpen the predictive value of other variables, even though class may predict poorly itself.

Discussion of these will be postponed beyond this paper.

#### The Nature of the Classification Rule:

By this is meant a rule which will assign any individual with information on a set of variables (the X's) to one of several mutually exclusive classes. The forms such classification rules might take are quite varied, and some choice must be made before any method to maximise relevance can "get off the ground." One particular form of rule will be outlined here, and it will be used as the basis of a procedure to seek maximum relevance for its classes.

Let us specify one location in X-space for each group. Each location would of course be defined by as many numbers (co-ordinates) as there are X-variables. The class membership of any case can be determined simply by seeing which location is nearest to that case. We will use Euclidean distance as a measure. Parenthetically, by various transformations of the X's beforehand, Euclidean distance can be made to reflect almost any desired kind of similarity.

Each location need not be close in an absolute sense to the members of its group, but it would of course always lie in the same region. The boundaries in X-space defined by this classification rule would be straight lines, planes, or hyperplanes; they are segments of the perpendicular bisectors between the various locations. The regions of X-space produced would sometimes be bounded on all sides, sometimes not. Regions would not overlap, and each region would always be convex. Such a system, while logically simple, still permits a fair degree of flexibility in the size and shape of possible regions which define classes. A set of locations, once defined, can then be used later to classify other cases with the same X- measurements -- i.e. the system can be "cross-validated" -- and its relevance to various other outside variables may be seen. The mean Y value of each class can be used to make absolute predictions of Y for new cases.

Such a system is really a species of nonlinear prediction function, and it may be evaluated in comparison with other sorts of predictions; for instance linear regression, quadratic regression, etc. Thus not only may the relevance of various systems of this sort be compared, but the efficiency of the whole classification approach may be compared to the efficiency of other ways of making predictions.

Such comparisons would probably be a very healthy thing for the field of cluster analysis. They would give us a more realistic idea of what we are and are not accomplishing. Predicting via discrete classes has both potential advantages and disadvantages when compared to other ways of using information to make specific forecasts.

The nature of the relation of Y to the X's obviously affects the relative success of using classes vs. the more usual kinds of algebraic functions to predict Y. This is seen most clearly when there is only one X-variable so that relations may be drawn on an ordinary graph. If the relation consisted of



several horizontal line-segments with sharp discontinuities in-between, then Y could be predicted perfectly via discrete classes. On the other hand, if it were a single slanted straight line, then predicting Y by a linear function of X would be better than using any finite number of discrete classes. Smooth but curvilinear relations could give various results, depending on the degree of non-linearity, complexity of the curve, number of classes used, etc.

Non-linear algebraic functions of the X's -- exponential terms, etc. -- would sometimes do better than either classes or linear terms. However, there is no real limit to the possible variety and complexity of non-linear functions but there usually is a limit to the size of the data sample.

We shall now return to the problem of how best predict Y via classes defined on the X's.

#### The Approach:

To make the problem of maximizing the relevance of such a classification rule a tractable one, we need to limit the number of classes. If there are as many classes as cases in the data sample, then we have the Nearest-Neighbor prediction rule of Fix and Hodges (8), which again indicates the overlap between this kind of cluster analysis and prediction theory. But when a data sample contains hundreds or thousands of cases, that procedure becomes an increasingly unwieldy classification or prediction system; all the work is in using the system, rather than in developing it. One would expect cross-validation performance to be poor, and such a system itself is not at all interpretable or easy to describe and communicate. The optimum number of classes would depend upon the relative gain in explanation or prediction from additional classes versus the loss in the form of extra mechanical or conceptual effort required to use it.

For the present we shall dodge this problem by holding the number of classes constant, and the solving for maximum relevance. Of course, a variety of solutions with different numbers of classes could give an indication of the point of diminishing returns in any particular applied problem.

When N is over 30 or so, I am sure that finding the optimal classification into some given number of groups is beyond us for some time, just as is the simpler problem of finding the minimum-variance partition among the X's alone. However, as in the latter case, it may be possible to find some very good solutions -- with a fair degree of certainty -- even if not necessarily the best solution that we are accustomed to getting from least-squares methods.

On the minimum-variance problem, I found (9,10) that one way of getting a number of good solutions is to start with a number of poor solutions. For instance, some arbitrary or random partitions, or with results from another method susceptible to improvement. Then I applied various improvement algorithms, making changes only when there was a demonstrable gain from doing so. One of the surprising (to me) results of such a procedure was that the goodness of the final classification depended very little, and not at all systematically, upon the starting-point. The computer program very rarely got "hung up" upon a very poor solution, regardless of what it started from.

Hopefully this may eventually be true of the same approach applied on the present problem, even though the improvement algorithms will differ.

Starting classes as well as final solution classes will be defined in terms of point locations (which we shall call P's), one for each group. A convenient source of "reasonable" P's -- that is, ones which produce groups all with at least one member -- is the data sample itself. If the P's are simply set equal to the X-coordinates of certain cases (perhaps chosen at random), then we have a starting classification with the desired geometric properties and necessarily with at least one member in each class. Alternatively, prior information might be used to define the starting P's directly. Another starting point would be to perform a minimum-variance type of clustering on the X-variables alone. If a single starting point is desired this might well be a very sensible one. However, if a number of widely different starting points are desired, they could be achieved most easily by a process of generating random selections of case numbers and setting the defining locations equal to the X's values of these cases.

#### The Improvement Algorithm:

= Central to the whole process is a fairly economical way of exploring a limited set of possible changes in the definition of classes, evaluating such changes, and making them whenever the between-class variance of Y is increased.

= Given the chosen way of defining classes, an obvious way of changing the classification system is to move a one or more of the locations. But in what direction, and how far?? In multivariate data there are an infinity of different directions that might be tried, and in each of these there's no guarantee that, as a location is moved, the variance of Y accounted for will change in a simple way. On the contrary, given the discrete nature of data points, the occurrence of multiple minima is to be expected.

On the question of how far to move, fortunately there are some natural limits. Moving in any direction, there is a limit beyond which a defining point will cause some group (usually its own) to have no members at all. Moving past this limit would be pointless, since it changes the nature of the problem by reducing the number of classes. In some directions there will be a closer "natural" limit, because continued movement would lead to crossing the old boundary of the region defined by its original position. Making this boundary an additional limit for a move would have the effect of allowing only moderate changes in the classification system during any one cycle of the iteration.

Now back to the question of direction for the move. Imagine a region with large Y variance, and in which Y is somewhat systematically related to the X's. Such a region is a good candidate for being "purified" -- made more homogeneous on Y -- which would improve the whole system. If the relation of Y to the X's is not too complex within each region, then fitting a regression plane to it would be a fairly accurate summary of this relation. (In practice, the slope coefficients for the regression planes should be computed in a "stepwise" fashion, so that there will be a solution even when a group has fewer points than there are X-variables.) This regression plane might provide a promising direction for a future move of the location. If it is moved "up" the plane (i.e. in the direction where high-Y values are found), then it will often change

the boundaries in such a way as to leave out some of the low-Y cases behind. If the location is moved "down" the plane (toward low-Y cases) then it will often act so as to leave out the high-Y cases behind it. In some cases, moving the location either direction may improve the classification system. The only feasible way to survey the effects of moving locations seems to be to evaluate the whole classification system at a number of specific moves -- perhaps at equal intervals between the permitted limits -- and choose the best location found.

The simplest way to move would be to move one group at a time, and proceed through the groups sequentially. If any group is improved by a move, then it may be worthwhile to cycle back and go through the groups again, because the move of a single group may change the boundaries and thus the membership of all the other groups. The process terminates when an entire cycle of groups has been gone through with no further improvement.

#### The "Y-GROUPS" Computer Program:

The two flow charts show the structure of the program and of the central "MOVE" Process. The user's description will give an idea of the various options and limitations of the present program. It was written in Fortran 4, and has performed on the IBM 360/75 at the Health Sciences Computing Facility at UCLA. It is still under development but copies of the program may be obtained from the writer.

#### Performance of "Y-GROUPS" on Test Problems:

Artificial data in known configurations involving 1, 2, and 4 X-variables were generated and used to test the algorithm, which proceeded from five random starting partitions on each problem. The results were excellent when no "noise" X-variables were present, but somewhat uneven over the several (random) starts when half or more of the X-variables were unrelated to Y in any fashion. All problems involved nonlinear relations of the X's to Y, and in all even the worst "Y-GROUPS" solutions were superior to linear regression in terms of the amount of Y variance accounted for.

#### Discussion:

While the present "Y-GROUPS" Program handled the test problems fairly well, it is only a first step toward a dependable method to establish maximally relevant classes. Some detailed analysis of its occasional "mistakes" on the several test problems should lead to improvements in the algorithm or perhaps even to a better general approach.

While it is, to my knowledge, the only cluster analysis method directed at relevance to outside measures, a number of developments in other fields have goals that are quite similar. For example, the "learning machines" of Hunt (12) and others establish a very different kind of classification rule in X- space -- one composed of logical "and's" & "or's", and make it successively more complicated until it achieves perfect relevance within the data sample. The many configurational scoring and pattern prediction methods of psychologists are

also pertinent, though we do not think of them as defining classification systems. However, again they would form a very complex and logically "messy" typologies, and their cross-validation performance has usually been quite disappointing. Yet another heading under which related work is being done is that of "pattern recognition" (13). And the problem of finding the optimum way to stratify a sample (14) is essentially similar, although I believe it has only been handled neatly only in situations with one variable.

Thus, while there is not yet much being directed at making simple clusters, classes, or types relevant to outside measures, there has been a great deal of effort and accomplishment at making other functions of multivariate data as relevant as possible to a wide variety of outside variables. There is almost certainly much to be learned from this work.

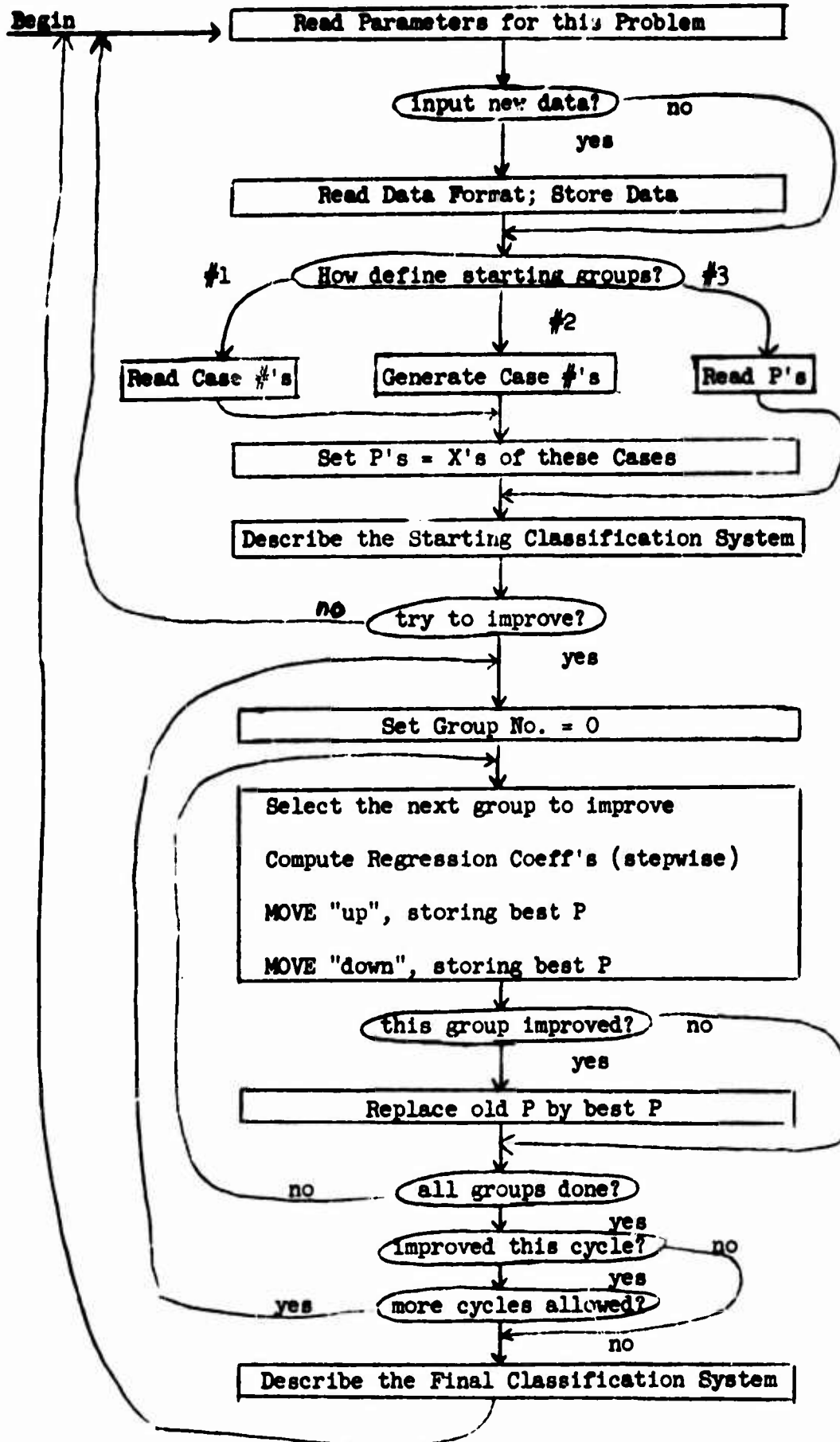
And even with an imperfect method in hand, I think there is also much that could be learned by applying it to empirical data which have been analyzed in enough other ways so as not to be a completely unknown quantity. I hope to do some of this myself in the near future, with some MMPI data in which relations are alleged to be distinctly nonlinear. Looking at the same body of data from several different perspectives will certainly help us understand clustering methods better than we do, and this should eventually help the empirical investigator come to more reasonable conclusions about what his data mean.

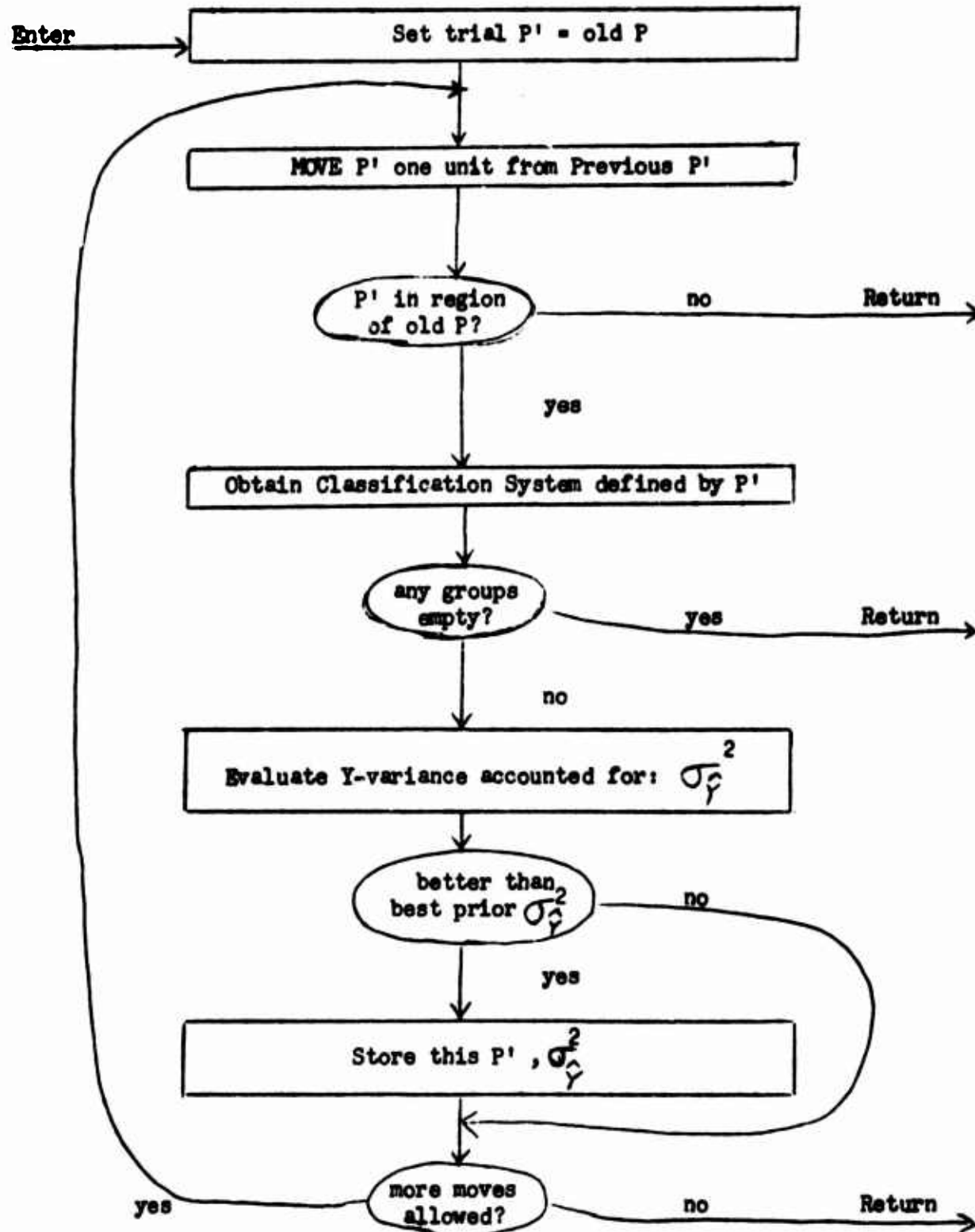
# REFERENCES

1. Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965. Under the Auspices of the American Psychiatric Association and the Psychopharmacology Service Center of The National Institute of Mental Health.
2. Lorr, Maurice. "Syndrome Based Psychiatric Types." Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965.
3. Overall, John E., and Hollister, Leo E. "Studies of Quantitative Approaches to Psychiatric Classification." Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965.
4. Katz, Martin. "A Phenomenological Typology of Schizophrenia." Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965.
5. Stein, Morris, and Neulinger, John. "A Typology of Self-Descriptions." Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965.
6. Mattson, Nils, and Gerard, Ralph. "Typology of Schizophrenia Based on Multi-disciplinary Observational Vectors." Conference on the Role and Methodology of Classification in Psychiatry and Psychopathology. Washington, D.C., November 19-21, 1965.
7. MacQueen, J. "The Classification Problem", Western Management Science Institute Working Paper No. 5, February, 1962.
8. Fix, Evelyn, and Hodges, J. L. Jr. "Discriminatory Analysis. Non-parametric Discrimination: Consistency Properties." Project No. 21-49-004, Report #4 (first published Feb. 1951). School of Aviation Medicine, Randolph AFB, Texas.
9. Forgy, Edward. "Cluster analysis of multivariate data: Efficiency vs. interpretability of Classifications." W N A R meetings, University of California, Riverside, June 22-23, 1965. (See abstract, Biometrics, 21(3), p. 768).
10. Forgy, Edward W., and Jennrich, Robert I. "Improving Classification Systems for Multivariate Observations." 1966. Manuscript.
11. MacQueen, James B. "Some Methods for Classification and Analysis of Multivariate Observations." Fifth Berkeley Symposium in Probability and Statistics. (In press)

12. Hunt, E., Marin, J. & Stone, P. Experiments in Induction, New York: Academic Press, 1966.
13. Sebestyen, G. Decision Making Processes in Pattern Recognition, New York: MacMillan, 1962.
14. Dalenius, Tore "The Problem of Optimum Stratification", Chapter 7, in Sampling in Sweden, Almquist & Wiksell, Stockholm, 1957.

Flow Chart for "Y-GROUPS" Program



Flow Chart for "MOVE" Subroutine in "Y-GROUPS" Program

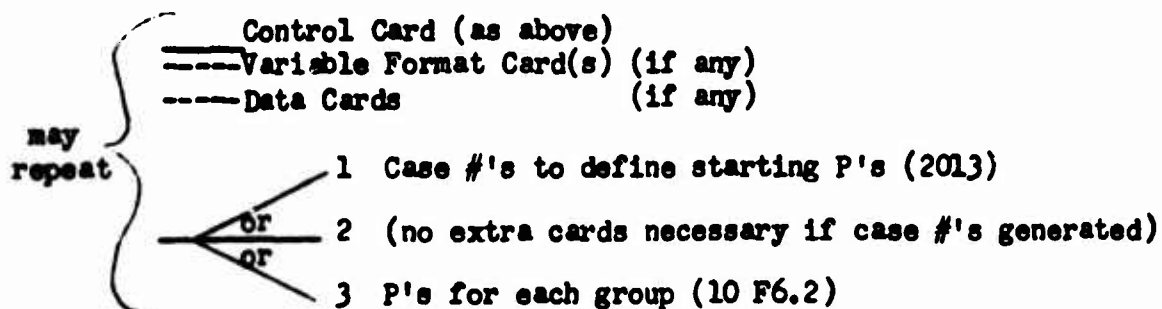


USER DIRECTIONS FOR "Y-GROUPS"

E. W. Forgy, UCLA  
November 1966

Problem Control Card:

column	symbol in program	control parameter	limits
1,2	NV	# Variables, <u>incl. Y</u>	2-31
3,4	KY	Variable # of Y	1-31
5,6,7,8	NC	# Cases	2-500
9,10	NG	# Groups	2-20
11,12,13	NST	# Starts	1-999
14	IST	Type of start: 1 = read in case #'s 2 = generate case #'s 3 = read in P's	1,2,3
15	NVFC	# Variable Format Cards for Data (F-type) (if 0, data from best problem is reprocessed)	0,1-5
16,17	NTAPE	Data Tape #, if tape input is used	-
18,23	DMV	Distance for each trial move of P: (punch decimal point)	-
24,25,26	MXNCYC	Max. # Cycles (if 0, starting classif. only is described)	0,1-999
27,28,29	MXNMV	Max. # Moves of a P in one direction	1-999
30	IDSC	1 - describe classes <u>after each move</u> 2 - describe classes <u>after each cycle</u> (otherwise, starting and final classes described)	0,1,2
31-80	PROB	alphanumeric label for this problem	-

Card Sequence: (after program)

## Investigation of Reliability of Different Profile Similarity Indices

John E. Overall  
The University of Texas Medical Branch

Two distinct problems in the methodology of cluster analysis have become apparent in this conference. The first problem concerns what is clustered and the second concerns how. In general, most profile clustering techniques involve first the computing of a matrix of similarity indices among all possible pairs of profiles and secondly the analysis of this matrix to identify subsets or clusters characterized by relative homogeneity within cluster and relative independence between clusters. Most discussion has been directed at the problem of how to identify homogeneous clusters, with the tacit recognition that most methods can be applied to a variety of different kinds of profile similarity measures; however, important questions exist concerning the meaningfulness and psychometric properties of the profile similarity indices that provide a basis for clustering. If clustering is to be meaningful and valid, reliability must be an important consideration in choice of the profile similarity index to be used.

The distance between two multivariate profiles can be considered to be a measurement statistic. This paper is concerned with an empirical investigation of distance function reliabilities, or more specifically with the consistencies between interprofile distances derived from rating profiles provided by two independent observers for the same samples of subjects. Considering interprofile distance to be a measurement, the investigation is concerned with simple interrater reliabilities of distances computed in different ways.

Several different methods for computing interprofile similarities which can be conceived as representing Pythagorean distances in Euclidean geometric space have been proposed in the literature. They differ in manner of defining coordinate axes and in the extent to which properties of the geometric model are identified with properties of the measurements. In order for a geometric system to serve as a model, certain points of coincidence need be established between the abstract geometric model and the measurement domain it is supposed to represent.

With regard to univariate measurement scaling, Stevens (1952) has discussed the problem of establishing points of coincidence, isomorphism, between real world and model. Different levels of measurement - nominal, ordinal, interval and ratio-scales - represent different degrees of correspondence between abstract number system and the real world it is supposed to represent. In the measurement of distances between multivariate profiles, a similar problem exists. On the one hand is the abstract Pythagorean model with mutually orthogonal reference axes of unit scale interval; on the other hand is a multivariate measurement domain. If the geometric system is to serve as a model, certain points of coincidence

# Investigation of Reliability of Different Profile Similarity Indices

## John E. Overall

must be established. It is clearly nonsense to say that "we may legitimately employ the Pythagorean model to calculate interpoint distances without concern for the degree of correlation among profile elements provided it is assumed that our coordinate axes are mutually orthogonal" (Heermann, 1965). The coordinate axes in the model are orthogonal, and no assumption is involved there. To what properties or characteristics of nature does the "assumption" apply?

The lowest level of correspondence, perhaps best conceived as analogous to nominal scaling, involves associating with each orthogonal axis of the model a single measurement variable, without regard for correlations among measurements or comparability of scale units. At this lowest level of isomorphism, the angles between the reference axes and the units of distance along the axes have no meaning with regard to statistical properties of the measurements. While it is true that we can use the simple Pythagorean formula to calculate interprofile distances without necessity for associating geometric angles and axis lengths with any properties of the data, the meaningfulness of such calculations appears questionable (Overall, 1964).

The degree of correspondence between abstract geometric model and measurement domain can be increased by equating statistical properties of the data with orthogonality of reference axes and with units of length of the geometric axes. For example, reference axes can be associated with statistically independent, equal-variance transformations of the original data. The orthogonal transformations can be obtained in a variety of ways, and different numbers of transformed orthogonal variates can be employed in computing interprofile distances. The variations in nature and number of transformed variates provide different distance measures which have been the subject of this investigation. Is orthogonal transformation useful? If so, what kind and how many transformed variates should one use? Reliability is one criterion to consider in evaluating answers to these questions.

The first distance index of interest is the simple  $d^2$  statistic (Cronbach and Gleser, 1953).

$$(1) \quad d^2 = d_1^2 + d_2^2 + \dots + d_p^2 = \underline{d}' \underline{d}$$

Recognizing that profile elements may be correlated to differing and unknown extent and that units of measurement may lack comparability, an orthogonal transformation of the original  $p$  correlated measurements may be sought to yield a new set of  $p$  uncorrelated, equal-variance transformations of the  $p$  correlated profile variates. Such transformation can be obtained using the inverse covariance matrix. The distance function computed from the transformed variates will be called a Mahalanobis - type  $D^2$ . (Since the Mahalanobis distance between groups is something quite different from this simple interprofile distance, we are probably doing the late Professor Mahalanobis a disservice in using this terminology).

$$(2) \quad D^2 = \underline{d}' \underline{C}^{-1} \underline{d}$$

# Investigation of Reliability of Different Profile Similarity Indices John E. Overall

Finally, another orthogonal transformation of the original profile elements based on factor analysis of the correlation (covariance) matrix has been suggested by the present investigator as having certain appeal. If the total variance is employed in the principal diagonal of the matrix of intercorrelations among profile elements, factor variates which have statistical properties of orthogonality and equal variance can be obtained (Overall, 1962). Distances between profiles can be computed using the Pythagorean model such that angles between reference axes and unit axis lengths have meaning in terms of the statistical properties of the transformed variates. In addition, the factor variates may have meaningful psychological interpretation, increased measurement reliability and other desirable properties.

$$D_F^2 = d_{F_1}^2 + d_{F_2}^2 + \dots + d_{F_p}^2 = \underline{d}' W W' \underline{d}$$

where  $W = C^{-1}F$  (for factors extracted from covariance matrix), or where  $W = C^{-1} V F$  in which  $V$  is a diagonal matrix containing test standard deviations (for factors extracted from a correlation matrix).

## Relationships between the Three Indices of Profile Similarity.

The simple  $d^2$  statistic (1) is a special case of the Mahalanobis-type  $D^2$  statistic (2). If it can be assumed that profile elements are uncorrelated and have equal variances, the inverse covariance matrix  $C^{-1}$  in equation 2 will be a diagonal matrix proportional to an identity matrix by a scalar constant.

$$D^2 = \underline{d}' C^{-1} \underline{d} = \underline{d}' I \underline{d} = \underline{d}' \underline{d} = d^2$$

If, on the other hand, profile elements are not uncorrelated and variances are not equal, then the simple  $d^2$  statistic may be quite different from the transformed  $D^2$  statistic. As Cronbach and Gleser (1953) have pointed out, the failure to take into account profile-element correlations results in statistically orthogonal factors being weighted according to the extent of representation in the profile, while in the transformed  $D^2$  each orthogonal dimension is weighted equally.

The Mahalanobis-type  $D^2$  (equation 2) is a special case of the general factor space  $D_F^2$  (equation 3). If factoring is continued until  $p$  orthogonal factors have been extracted from the  $p$ -order covariance (correlation) matrix, the  $D_F^2$  computed from the  $p$  transformed orthogonal factor variates will be precisely the Mahalanobis-type  $D^2$  for the same profiles. (This equivalence will be illustrated only for the factoring of covariance matrix; however, it should be obvious that the same relationship holds for factors extracted from a correlation matrix when loadings are re-scaled through multiplying by test standard deviations.) When the covariance matrix is factored completely, it can be reproduced perfectly from the factor loadings.

$$C = F F'$$

## Investigation of Reliability of Different Profile Similarity Indices

John E. Overall

The factor score transformation matrix is obtained

$$W = C^{-1} F.$$

The factor space distance function  $D_F^2$  is computed by equation 3.

$$D_F^2 = \underline{d}' W W \underline{d} = \underline{d}' C^{-1} F F' C^{-1} \underline{d} = \underline{d}' C^{-1} \underline{d} = D^2$$

Thus, the Mahalanobis-type  $D^2$  is precisely equivalent to the factor space  $D_F^2$  in the special case where factoring has proceeded to extraction of all  $p$  factors. Where factoring is terminated after  $r < p$  factors have been extracted, the Mahalanobis-type  $D^2$  may be substantially different from the factor space  $D_F^2$ .

If we conceive that a matrix may contain only  $r < p$  reliable factors and that additional factors may represent only error variance, we have a basis for understanding the very considerable differences in reliability of results which will be reported to exist between the alternative approaches. If a matrix is factored completely and orthogonal factor variates are all scaled to equal variance, the effect will be to increase greatly the error involved in assessing profile similarities when, in fact, there are only a few true common factors and many small error factors (now stretched to unit length just like the true factors).

These results appear to mediate against my previous recommendation of the complete Mahalanobis-type  $D^2$ , not because it is unimportant to establish coincidence with geometric properties of orthogonality and equal unit coordinate axes, but because the coordinate axes need to be defined in terms of true, non-error factors. Since the Mahalanobis-type orthogonal transformation is equivalent to complete orthogonal factoring, an interesting question arises concerning how many transformed orthogonal variates should be used in computing interprofile distances.

### Empirical Study of Reliability of Distance Indices.

In psychiatric symptom ratings, the degree of agreement between two independent observers represents an important kind of reliability. Unless two observers can agree concerning the level of symptomatology present in each patient, there is little basis for confidence that the ratings represent true status of the patients. Where psychiatric rating profiles are used as a basis for clustering of patients with the hope of identifying naturally occurring homogeneous modal types, it is important to know the extent to which the same cluster results can be expected to result from ratings made by different observers. If the relative distances between patients differ widely from one independent observer to the next, one can have little confidence that the cluster results really represent fundamental types of patients.

The present investigation was undertaken on the assumption that some types of profile similarity indices may be more invariant (reliable) across different observers than others. The investigation involved

# Investigation of Reliability of Different Profile Similarity Indices

## John E. Overall

comparing interprofile distances derived from ratings by one observer with interprofile distances derived from ratings by another observer. This comparison was made for seven different measures of profile similarity, representing variations of the three basic models described above ( $d^2$ ,  $D^2$  and  $D_F^2$ ). The analyses were replicated across seven independent random samples of 20 patients - each sample yielding  $n(n-1)/2 = 190$  interprofile distances for ratings by each observer. The seven different distance indices for which interrater reliabilities were evaluated are shown in Table 1.

The first three series of analyses involved interprofile distances computed using only the information present in each sample of 20 profiles being analyzed. In the case of the simple  $d^2$  index, this is all information that can be used since no transformation of the original variates is imposed. With the Mahalanobis  $D^2$ , the original variates are transformed to a set of mathematical variates which are statistically orthogonal in some population or in some sample. When inter-profile distances are computed using only the information present in the sample, the transformed variates are statistically orthogonal within that sample. Such an orthogonal basis contains the sampling error present in the specific small-sample covariance matrix; hence, it may be different from one rater to the next. Variations in the covariance matrix, thus, contribute to variability of  $D^2$  results from one rater to the next, even within the same sample of patients. Using the factor space  $D_F^2$  model, the original variates are transformed to a set of  $r < p$  mathematical variates which are statistically orthogonal in the sample or population represented in the correlation (covariance) matrix which is factored. When the orthogonal factor variates are derived from analysis of the small-sample correlation (covariance) matrix involving only the cases for which inter-profile distances are being computed, the factor variates will be influenced by sampling variability in the covariances. Since the covariance matrix will differ from one rater to the next, some variability in distance function results may be introduced. On the other hand, common factors tend to be more stable than individual variables so that reliability may be increased.

The three types of distance indices were computed for all possible pairs of patients in the seven samples, first using ratings by one rater and then using ratings for the same patients made by another rater. The  $n(n-1)/2 = 190$  resulting paired distance indices in each sample were intercorrelated for the two raters. In this instance, the rank correlation coefficient was employed as a simple descriptive index of the relative similarities of distance indices computed from ratings by the two independent observers. (No assumptions concerning distributions of these coefficients were made.) The results of correlations between paired distance measures for the three types of indices are plotted in Figure 1 for the seven independent samples of patients.

The results indicate that the orderings of simple  $d^2$  indices were consistently most similar for the two independent raters. The factor space  $D_F^2$  results were less consistent from rater to rater when the factor variates were defined in terms of the individual sample ( $N=20$ ) correlation



# Investigation of Reliability of Different Profile Similarity Indices

## John E. Overall

matrices for each rater separately. Finally, the ordering of Mahalanobis-type  $D^2$  indices was almost entirely lacking in consistency from rater to rater when the individual sample ( $N=20$ ) covariance matrix was used in computing  $D^2$ . In fact, inter-rater correlations were equal to or less than zero in four out of the seven samples.

Figure 1

The results of this first series of analyses leads to the conclusion that the simple  $d^2$  index of profile similarity is significantly (7 out of 7) more invariant across raters than either the factor space  $D_F^2$  or the Mahalanobis-type  $D^2$  when only the information contained in the profiles being clustered is used. The results further suggest that the Mahalanobis-type  $D^2$  is entirely lacking in reliability across raters when the small-sample covariance matrix is employed in the calculations. As previously discussed, this is due to the fact that the Mahalanobis-type  $D^2$  is equivalent to factoring the correlation (covariance) matrix completely and then equating the variance of all factor variates, whether true common factors or error factors.

The next series of analyses was undertaken to evaluate the effect of increasing the stability of the covariance matrix used in Mahalanobis-type  $D^2$  calculations. A single stable covariance matrix based on a larger sample ( $N=280$ ) was computed, and the inverse of this covariance matrix was used in calculating inter-profile distances in all samples for both raters. This procedure is equivalent to transforming all rating profiles using a common transformation matrix. It is like factoring a stable population correlation (covariance) matrix completely, equating variances of all factor variates, and using these factor variate equations in transforming all ratings. The results of this procedure were correlated for the two independent raters in the seven samples. Results are presented in Figure 2. Use of the more stable common covariance matrix to obtain orthogonal transformation, rather than obtaining a separate transformation matrix for each rater, resulted in increased inter-rater reliability for the Mahalanobis-type  $D^2$  indices; however, the inter-rater reliability was still found to be quite low. For comparison, results obtained for the same data using the simple  $d^2$  index are reproduced in Figure 2. Even where a single orthogonal transformations for the Mahalanobis-type  $D^2$  calculations, the simple  $d^2$  index evidences considerably more stable results from one rater to the next. Again, this result is presumable due to the increased emphasis on error factors resulting from a transformation which is equivalent to total factoring of a matrix containing only four substantial principal factors and 12 roots less than unity.

Figure 2

A final series of analyses was undertaken to evaluate the inter-rater reliability of factor space  $D_F^2$  indices computed from factor variates derived from a single large sample ( $N=280$ ) correlation matrix. Factor

Investigation of Reliability of Different Profile Similarity Indices  
John E. Overall

score transformation vectors were computed for one, two and four principal factors of the common large sample correlation matrix. Inter-rater correlations of  $D_F^2$  values were computed within each sample of 20 cases using the same factor score transformation equations. Results are presented in Figure 3. For comparison, the simple  $d^2$  results are also reproduced in this figure.

---

Figure 3

---

While the consistency from sample to sample is not as pronounced, the general trend is for the factor space  $D_F^2$  coefficients to evidence greater invariance between raters than the simple  $d^2$  statistic. Where  $D_F^2$  based on the four principal factors corresponding to latent roots greater than unity were analyzed, the inter-rater consistency was higher than for the simple  $d^2$  statistic in six out of the seven independent samples. The average inter-rater consistency increased when  $D_F^2$  indices were based on only first two principal factors; and increased still more with use of only first principal factor; however, the variability from sample to sample increased as fewer factors formed the basis for  $D_F^2$  calculations. As has already been pointed out, the  $D_F^2$  statistic approaches the Mahalanobis-type  $D^2$  as the number of factors approaches the total number of profile components. For comparison  $D_F^2$  inter-rater correlations have been entered in Figure 3 also. 16

From these results it is concluded that the use of a stable orthogonal transformation representing only the non-error factors of a large-sample correlation matrix will tend to result in more reliable profile similarity indices, that there is generally an inverse relationship between number of factors used in defining the space and the reliability of distance indices, but that the simple  $d^2$  statistic compares favorably to the best profile similarity measures, as far as inter-rater consistency is concerned.



Table 1

Seven Indices Employed in Comparing Relative  
Invariance of Distances Computed from Rating Profiles Provided  
by Two Independent Observers

	Simple $d^2$	Mahalanobis $D^2$	Factor Space $D_F^2$
Covariance Matrix based on N=20	$d^2 = \underline{d}' \underline{d}$	$D^2 = \underline{d}' C^{-1} \underline{d}$	$D_{F_4}^2 = \underline{d}' W W' \underline{d}$
Covariance Matrix based on N=280		$D^2 = \underline{d}' C^{-1} \underline{d}$	$D_{F_4}^2 = \underline{d}' W W' \underline{d}$
Two principal factors			$D_{F_2}^2 = \underline{d}' W W' \underline{d}$
One principal factor			$D_{F_1}^2 = \underline{d}' W W' \underline{d}$

Simple  $d^2$ :

Covariance matrix not involved; all variables enter into distance calculations.

Mahalanobis  $D^2$ :

First series of analyses involved use of  $C^{-1}$  calculated from the particular sample of 20 cases for whom interprofile distances were calculated. Second series of analyses involved use of a constant  $C^{-1}$  based on larger sample of 280 cases.

Factor Space  $D_F^2$ :

First series of analyses involved first four principal factors of sample (N=20) correlation matrix. Second series of analyses involved first four principal factors of constant correlation matrix based on large sample of 280 cases. Third and fourth series involved two and one principal factors of large sample matrix.

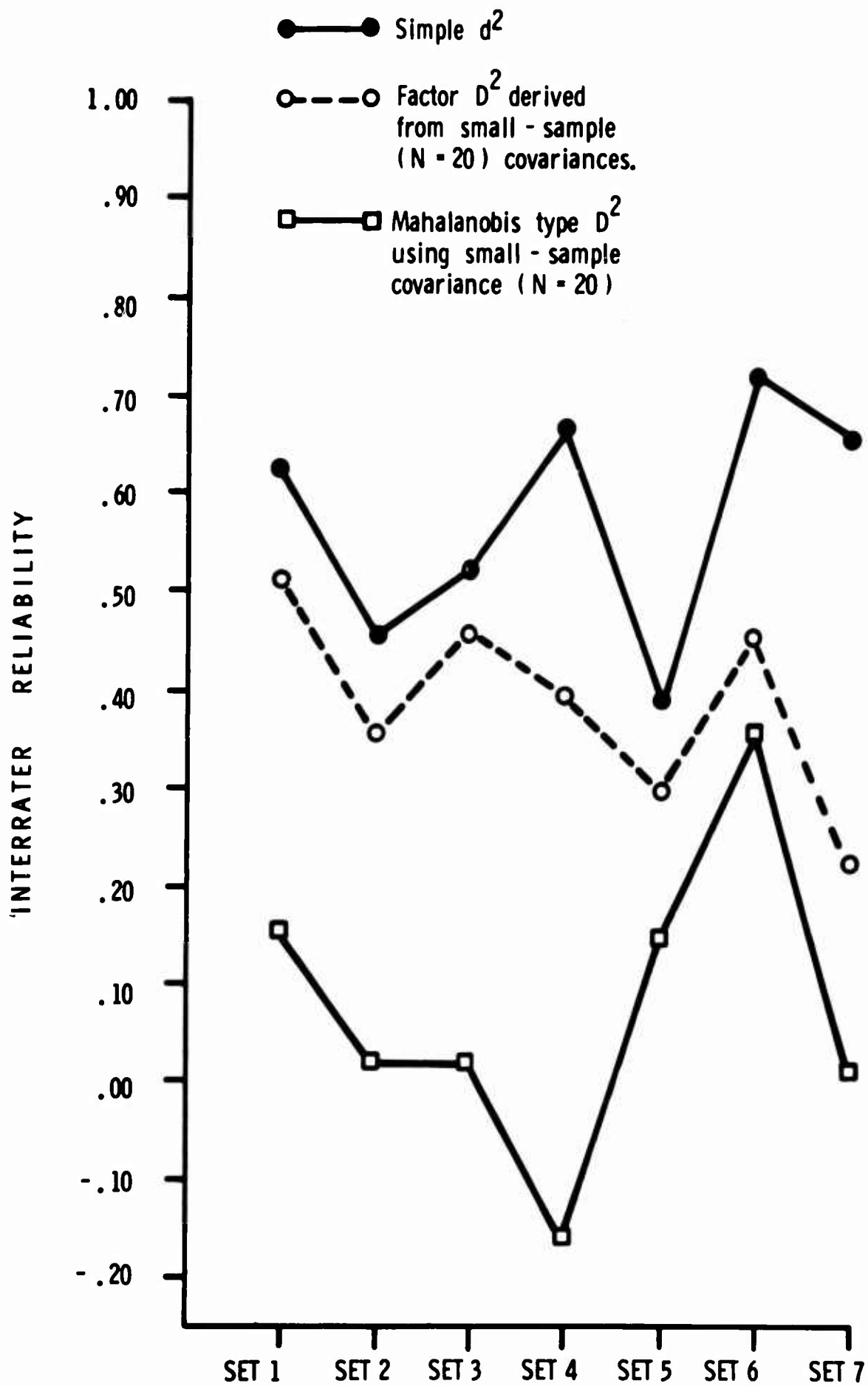


Figure 1

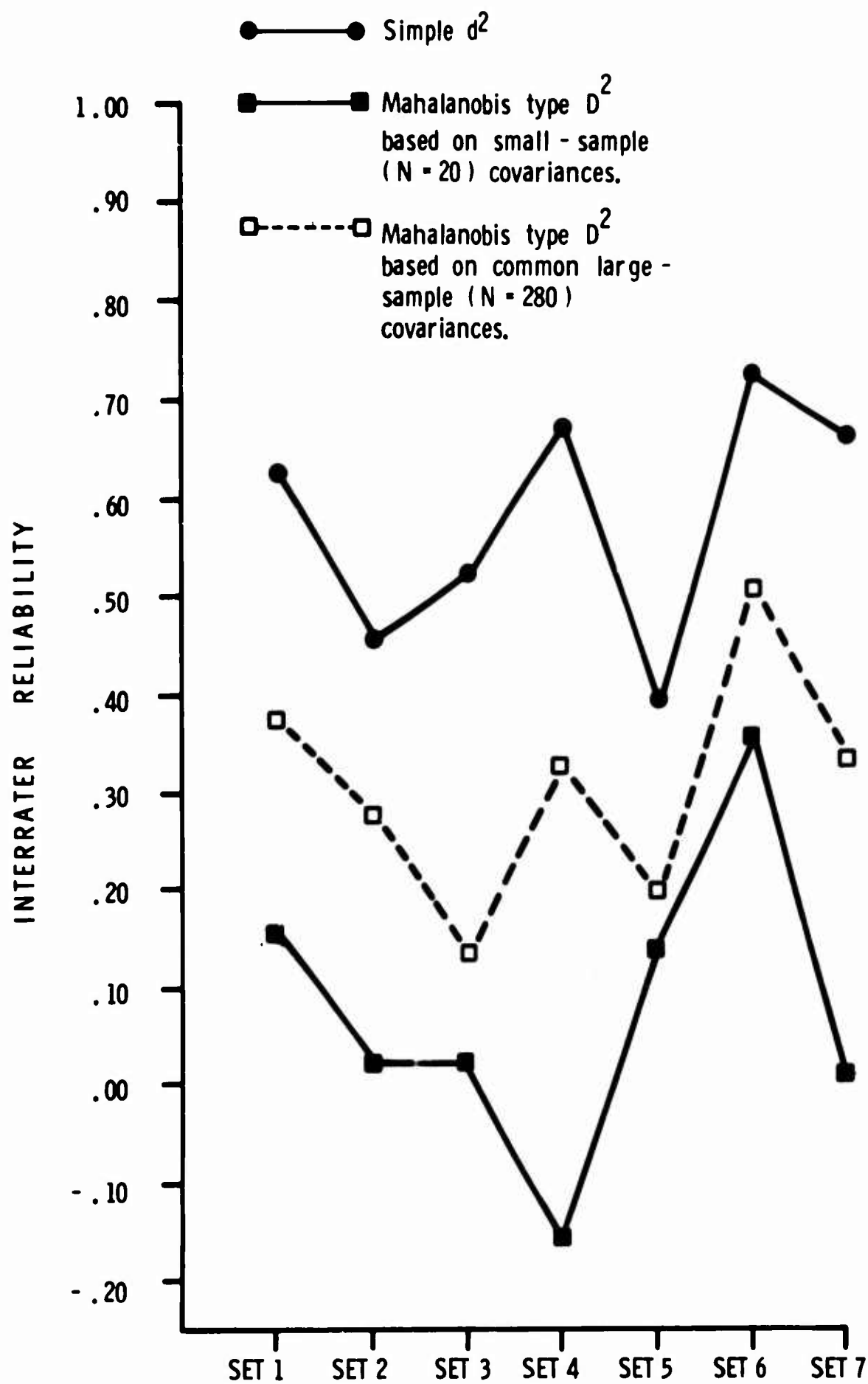


Figure 2

INTERRELATED RELIABILITY

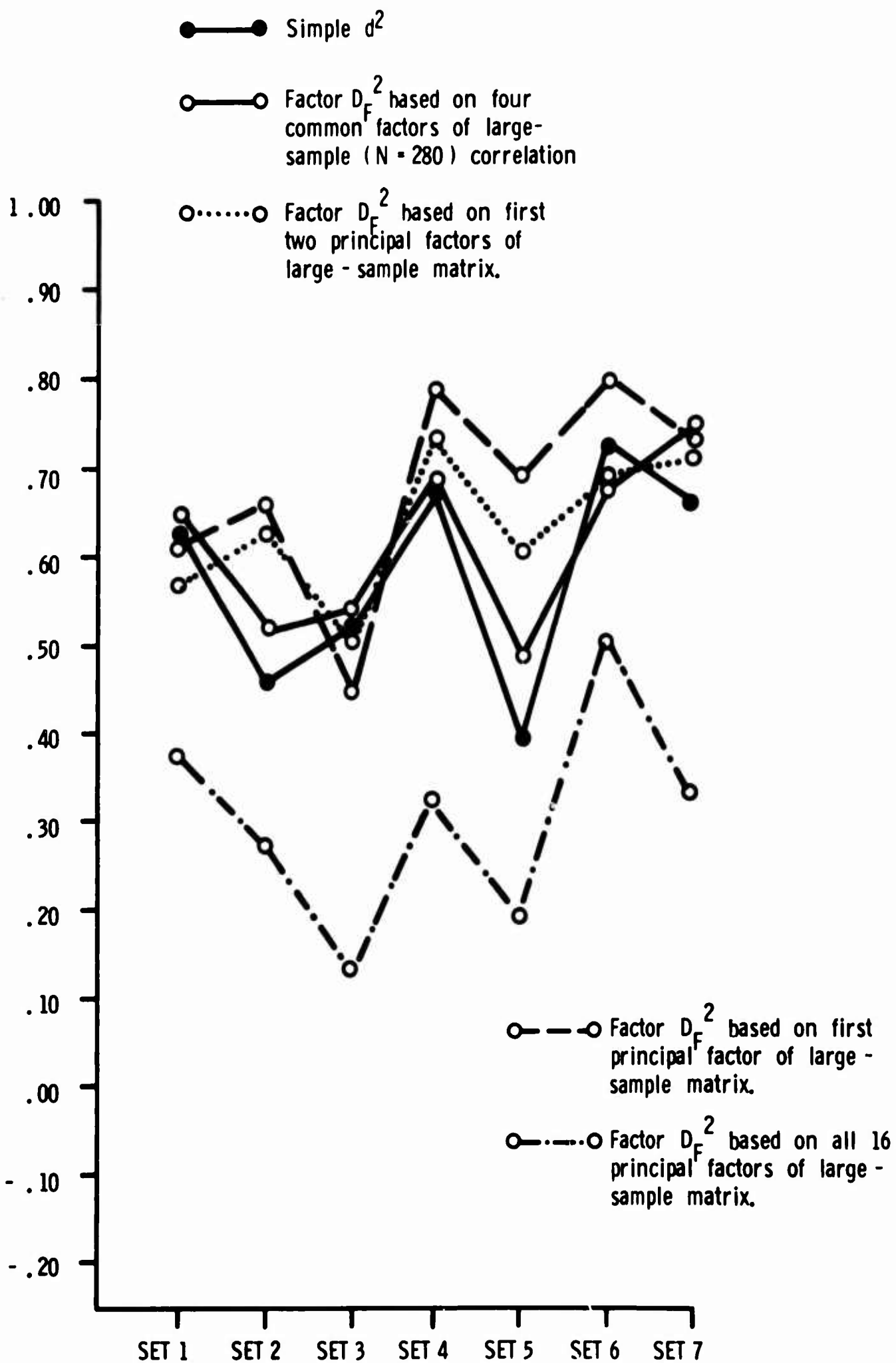


Figure 3